

# Leveraging LIME for Human-Understandable Explanations in Urban Computer Vision Models

Analyzing street view images for further urban liveability understanding

Master Thesis  
Bastiaan Bakker

# Leveraging LIME for Human- Understandable Explanations in Urban Computer Vision Models

Analyzing street view images for further urban  
liveability understanding

by

Bastiaan Bakker

Student Name	Student Number
Bastiaan Bakker	4605004

1st supervisor: S. van Cranenburgh  
2nd supervisor: A. Sepinoud  
3d supervisor: T. Perenboom  
Project Duration: February, 2024 - August, 2024  
Faculty: Faculty of Technology, Policy and Management, Delft

Cover: Street View Image of Rotterdam  
Style: TU Delft Report Style, with modifications by Bastiaan Bakker

# Preface

This research marks the end of my graduation project, which started on February 12, 2023, seven months ago already. Finishing this project also means the end of my student life in Delft. I began my bachelor's in Applied Mathematics in 2016 and never thought my master's thesis nearly eight years later would be about eXplainable AI, especially related to urban public spaces. Neither subject was covered in my bachelor's or master's programs, making this project even more educational, challenging, and fun for me.

While there were times I felt unsure about my knowledge of these topics, the journey was rewarding, especially since it involved my current city, Rotterdam. This connection made analyzing the city even more meaningful. Through this research, I learned a lot about eXplainable AI, computer vision models, public spaces, how municipalities work, urban liveability, and about myself and how to manage my own research.

This study would not have been possible without the help of several individuals. I want to thank Tommie Perenboom of the Municipality of Rotterdam for guiding me through the process and helping me understand how the municipality works, what it wants, and where to ask for help. I also thank the other colleagues at the Municipality of Rotterdam for their warm welcome and the conversations we've had.

I am also grateful to my supervisors, Sander van Cranenburgh and Sepinoud Azimi Rashti from the TU Delft, for always making time for me and giving valuable feedback on my work. Thanks also to the Data Analytics workgroup led by Sander van Cranenburgh, where I and other students worked on similar master's thesis subjects, giving us the chance to discuss these topics together.

Lastly, I want to express my heartfelt thanks to my family, friends, roommates, Kojaccers, and girlfriend for their unwavering support, not just over the past seven months, but throughout my entire eight years of study.

I hope you enjoy reading this research as much as I enjoyed working on it.

*Bastiaan Bakker  
Delft, August 2024*

# Summary

Human understanding and interpretability of AI models have become a growing problem in policy-making, particularly in urban planning. As AI models become more complex and opaque, decision-makers struggle to translate and explain their outcomes into human terms, leading to policy models that lack explainability and human-understandable outcomes. This research explores whether applying Explainable Artificial Intelligence (XAI) techniques, specifically Local Interpretable Model-agnostic Explanations (LIME), can enhance the interpretability of computer vision-enriched discrete choice models (CVDCMs) for street view images in urban analysis. CVDCMs analyze the liveability of areas using street view images, by analyzing the full image to determine its liveability score. This is unlike traditional urban models using object detection for street view images (relating liveability to a certain set of predefined objects in street views and measuring feature importances), thereby introducing a bias of the modeller in the object set choice. CVDCMs, in combination with XAI, can avoid this specific modeller bias, providing a more holistic view of human perception of urban liveability and improving the CVDCM's utility for policy advice by possibly providing new insights into urban liveability.

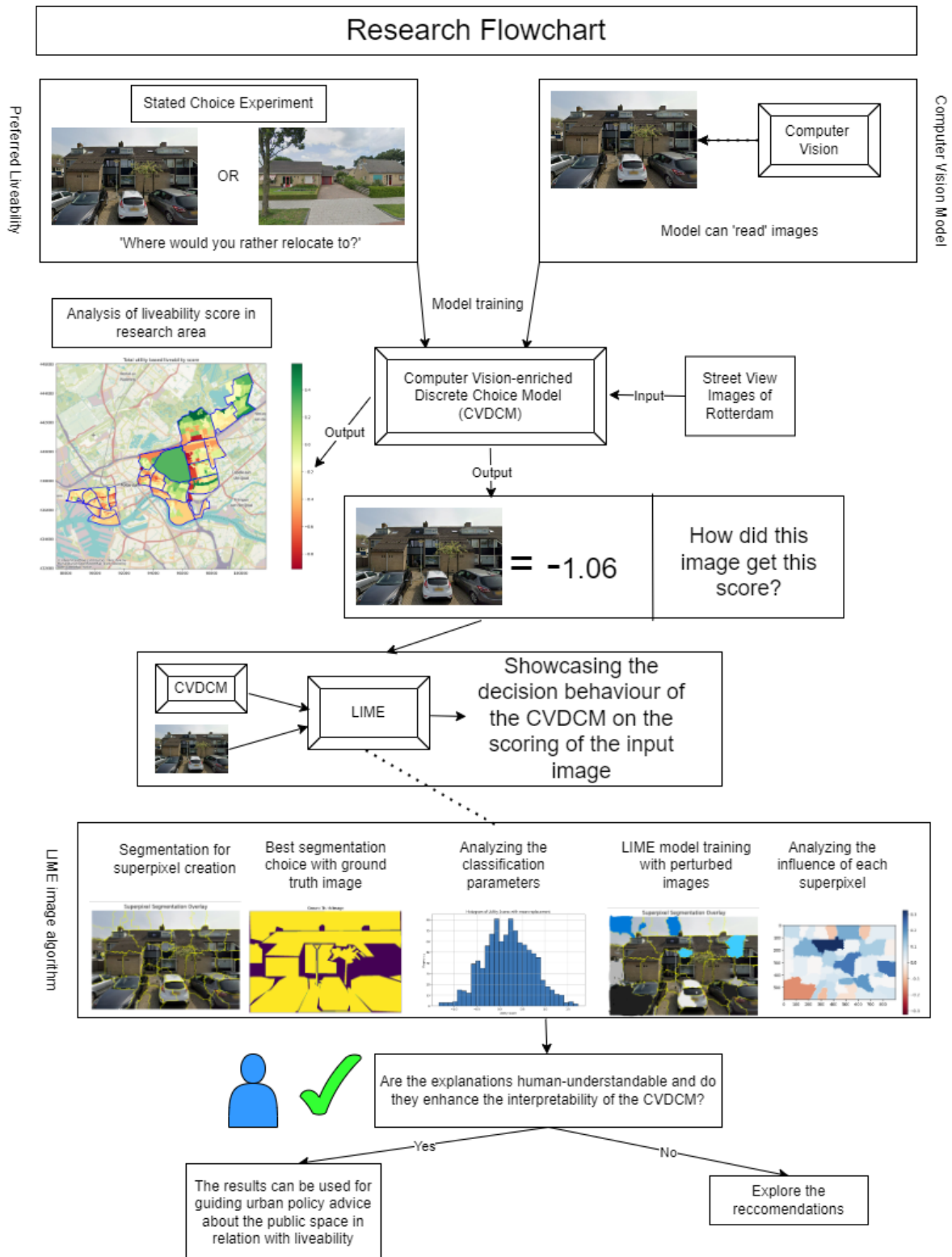
The methods employed in this research include applying CVDCMs to perform a face validity analysis with street view images to gain an initial understanding of the model's decision behaviour. LIME was then used to generate explanations for these model decisions by segmenting the images, perturbing these segments to create modified images, and analyzing the impact on the classification of the street views into 'good' or 'bad'. This approach highlights which 'parts' of street view images contribute most to preferred liveability, thus providing a more comprehensive view of human perception. LIME can showcase the complex decision behaviour of the CVDCM by considering the entire street view image as a possible explanation for a liveability score, without requiring a predefined set of factors. This contrasts with non-XAI methods, which leave users without insight into how the model arrives at its results. By not relying on prior object detection, LIME avoids modeller bias and allows for a more holistic view of human perception, enhancing the interpretability and transparency of CVDCMs. The LIME segmentations were evaluated with ground truth images and the LIME explanations were assessed with various metrics (Binary Classification Ratio, Coefficient of Variation, and Probability Distribution Uniformity Metric), ensuring robust and reliable interpretability of both the segmentations and the LIME explanations.

The study revealed several significant findings. The quality of LIME explanations is highly dependent on the segmentation process, as already also found in earlier research. Street view images, which contain a variety of complex objects, present challenges in achieving meaningful segmentation. While ground truth analysis can improve segmentation quality, it is labour-intensive and prone to inconsistencies. The lack of semantically meaningful segmentation results in non-human-understandable explanations of the model's outcomes, and thus does not increase the interpretability of CVDCMs. Additionally, the sampling process of LIME results in perturbed images that significantly differ from the original image. This impacts the utility-based liveability score and necessitates a unique classification threshold for each image in the deterministic classification, complicating comparisons and reducing explanation coherence. In the probabilistic classification, this is not necessary, but the distribution of the classification probabilities is therefore not adequate, resulting in scientifically less valid LIME explanations. Moreover, the computational intensity of the LIME methodology limits its scalability for analyzing large datasets necessary for comprehensive urban policy analysis. Therefore it is concluded that in the current applied methodology, the interpretability of the complex model is not increased and no human-understandable explanations are created. This research uniquely identifies the complications of using CVDCM with LIME image analysis and suggests that simpler object detection models combined with LIME tabular explanations could offer a valuable alternative.

In response to these findings, several recommendations are proposed to increase interpretability and

create human-understandable explanations. The use of street view-trained segmentation models for superpixel creation (via object detection) can provide human-understandable LIME explanations by providing semantically meaningful segmentations, which could also decrease the computational time of the LIME method, thereby introducing some bias but still enabling the analysis of the holistic view of human perception. Further research on the distribution of utility scores in the sampling process and experimenting with different parameters or distributions for the sampling process could improve the LIME explanations. Additionally, applying methods to integrate regression models directly within the LIME framework, rather than converting them to classification models, could avoid the pitfalls of threshold-based classification. These recommendations ensure that the explanations are more aligned with the process of urban policy advice creation, as they become logical and actionable through these improvements, enabling its use for municipal policy advice.

Figure 1 shows a simplified flowchart of the research, in order to provide more clarity and understanding for the reader on the steps taken in this research.



**Figure 1:** Simplified flowchart of the outline of the research

# Contents

<b>Preface</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Research question	2
1.2 Sub questions	3
<b>2 Literature Review</b>	<b>4</b>
2.1 Human Urban Perception	4
2.1.1 Preferred liveability	6
2.1.2 Perceived liveability	7
2.2 Computer vision in Urban research	8
2.2.1 Human perception overview	10
2.3 eXplainable AI in urban research	11
2.4 Research gap: XAI for urban research with computer vision models	13
<b>3 Methodology</b>	<b>15</b>
3.1 Data Retrieval for model usage	15
3.2 Computer-vision enriched discrete choice model	15
3.3 XAI techniques	20
3.4 Application of LIME for images	21
3.4.1 Superpixels	22
3.4.2 Sampling	22
3.4.3 Weights	22
3.4.4 Surrogate model	23
3.5 Segmentation metrics	23
3.5.1 Ground truth images	24
3.5.2 Segmentation quality metrics	24
3.6 LIME metrics	26
3.6.1 Binary class ratio	26
3.6.2 Probability Distribution Uniformity Metric	26
3.6.3 Coefficient of covariance	26
<b>4 Implementation</b>	<b>28</b>
4.1 Segmentation algorithm choice	28
4.2 Classification problem	32
4.3 Perturbation utility score changes	34
4.4 LIME Explanation visualization	36
4.4.1 Most important superpixels	36
4.4.2 Hide function	36
4.4.3 Heatmap	36
4.5 Interpretation of LIME	36
4.6 Implementation pipeline	37
<b>5 System and Actor analysis of human perception modelling at the municipality of Rotterdam</b>	<b>38</b>
5.1 System analysis	38
5.2 Actor analysis	39
5.2.1 Power-Interest Matrix	41
5.3 Interactions and Dependencies	44
5.4 Expert Analysis	44
5.5 Conclusion	45

<b>6</b>	<b>Face validity analysis of Rotterdam's utility-based liveability score</b>	<b>46</b>
6.1	Research location	46
6.2	Utility-based liveability scores of selected areas	49
6.2.1	Delftshaven	57
6.2.2	Prins-Alexander	61
6.2.3	Kralingen-Crooswijk	65
6.3	Face validity conclusions	69
6.4	Result validation	70
6.4.1	Colour influence	75
6.5	Validation conclusion	75
<b>7</b>	<b>Application of LIME: case studies</b>	<b>77</b>
7.1	Case study 1	77
7.2	Case study 2	85
7.3	Case study 3	88
7.4	Case study 4	90
7.5	Case study 5	91
7.6	Case study 6	93
7.7	Case study 7	95
7.8	Application of LIME conclusion	98
<b>8</b>	<b>Discussion</b>	<b>100</b>
8.1	Sampling Process and Utility Score Distribution	100
8.2	Complexities of Street View Images	100
8.3	Need for prior object detection	101
8.4	Subjective nature of LIME	101
8.5	Classification problem	101
8.6	Time costs	102
8.7	Comparison with Traditional Object Detection Models	102
8.7.1	Traditional object detection models with tabular LIME explanations	102
8.7.2	Other image XAI techniques	103
<b>9</b>	<b>Conclusion</b>	<b>104</b>
	<b>References</b>	<b>106</b>
<b>A</b>	<b>Globally known liveability frameworks</b>	<b>113</b>
A.1	Leefbarometer	114
<b>B</b>	<b>Actor Analysis Public Space Perception Municipality of Rotterdam</b>	<b>116</b>
B.1	Cluster Stadsbeheer	116
B.1.1	Afdeling Schone en Circulaire Stad	116
B.1.2	Afdeling Toezicht, Handhaving, Parkeren en Markten	116
B.1.3	Afdeling Openbare Werken	117
B.1.4	Afdeling Openbare Werken: Stedelijk Beheer	117
B.1.5	Afdeling Openbare Werken: Stedelijk Beheer: AMOR	117
B.1.6	Afdeling Openbare Werken: Stedelijk Beheer: CROW	117
B.1.7	Afdeling Openbare Werken: Gebiedsbeheer	117
B.2	Cluster Maatschappelijke ontwikkeling	117
B.3	Cluster Diensverlening	118
B.3.1	MELDR App	118
B.3.2	Wijken, Participatie, Stadsarchief	118
B.3.3	Wijken, Participatie, Stadsarchie: Wijkregisseurs	118
B.3.4	Wijken, Participatie, Stadsarchie: Wijkhubs	118
B.3.5	Wijken, Participatie, Stadsarchie: Wijk Aan Zet	118
B.4	Cluster Stadsontwikkeling	118
B.4.1	Afdeling Stedelijk Inrichting	119
B.4.2	Afdeling Gebiedsontwikkeling	119
B.4.3	Afdeling Economie en Duurzaamheid	119



---

B.5	Afdeling Onderzoek en Business Intelligence (OBI) . . . . .	119
B.5.1	Onderzoeksteam . . . . .	120
B.5.2	Advanced Analytics . . . . .	120
B.6	Wijkraad . . . . .	120
B.7	Gemeenteraad, het college burgemeester en wethouders . . . . .	120
B.8	Dutch National Government . . . . .	120
B.9	European Commission . . . . .	120
B.10	Citizens of Rotterdam . . . . .	120
<b>C</b>	<b>Utility based liveability analysis of all selected areas</b>	<b>121</b>
<b>D</b>	<b>Additional neighbourhood utility-based liveability plots</b>	<b>123</b>
<b>E</b>	<b>Prins Alexander en Kralingen Crooswijk mapped together</b>	<b>127</b>
<b>F</b>	<b>Quantile plots of utility-based liveability score for Rotterdam</b>	<b>129</b>
<b>G</b>	<b>Street view images analysis of neighbourhoods</b>	<b>132</b>
G.1	Witte dorp . . . . .	132
G.2	Bospolder . . . . .	133
G.3	Tussendijken . . . . .	134
G.4	Nesselande . . . . .	135
G.5	Het Lage Land . . . . .	136
G.6	Kralingseveer . . . . .	137
G.7	Kralingse Bos . . . . .	138
G.8	Kralingen Oost . . . . .	139
G.9	Kralingseveer . . . . .	140
<b>H</b>	<b>Additional analyzed segmentation algorithms</b>	<b>142</b>
<b>I</b>	<b>Segmentations with Quickshift, Felzenszwalb and SLIC for the example image</b>	<b>144</b>
<b>J</b>	<b>LIME case studies</b>	<b>151</b>
J.1	Case study 2 . . . . .	151
J.2	Case study 3 . . . . .	152
J.3	Case study 4 . . . . .	154
J.4	Case study 5 . . . . .	155
J.5	Case study 6 . . . . .	157
J.6	Case study 7 . . . . .	158

# 1

## Introduction

Recent advances in machine learning have positioned Artificial Intelligence (AI) as a key tool in assisting human decision-making across various domains, including policy advice in urban research (Meske et al. [92]). However, as AI systems become more advanced, allowing them to model more complex situations, they become more opaque. Opacity is when the model's decision processes and outcomes become difficult to interpret for users (Berente et al. [9]). This opacity creates the danger of users "blindly" trusting the outcomes without understanding the decision-making processes behind the model's results. Explainable AI (XAI) emerges as a solution to this challenge by generating human-understandable explanations that help users comprehend AI outcomes (Arrieta et al. [6]). XAI aims to equip users to explore AI decisions, evaluate their validity, and learn from them (Meske et al. [92]). While XAI can improve trust and performance, the effectiveness depends on meeting users' specific interpretability needs (Hamm et al. [48]).

In urban research, the advancement and application of computer vision (CV) models offer a promising approach by analyzing street view images to provide assessments of urban environments (Yigitcanlar et al. [143]). In this regard, computer vision-enriched discrete choice models (CVDCM) have emerged as a novel method for urban liveability research (Cranenburgh and Garrido-Valenzuela [23], Son et al. [123]). Here, urban liveability refers to individuals' subjective perceptions or evaluations of the quality of life in a particular place. Leidelmeijer and Kamp [70] defines it as the degree to which the urban environment aligns with the requirements and desires expressed by its citizens. The model from Cranenburgh and Garrido-Valenzuela [23] is used in this research, which computes liveability scores with street view images as input, and this model is created in cooperation with the TU Delft and the municipality of Rotterdam. This model is complex and inherently not interpretable. Because of this, a case study with the application of XAI to this model on the analysis of street view images of Rotterdam is performed. Due to their complexity, this CVDCM poses the possibility of uncovering the complex relations between street features in the urban environment. However, this complexity also poses significant challenges to their interpretability and creating human-understandable explanations of their results. Although these models can potentially inform policy advice, the lack of transparency in their decision-making processes could lead to misleading advice. Therefore, the interpretability and explainability of these models are vital. Next to this, providing humanly understandable explanations of how the decision behaviour of these complex models works.

To address these challenges, this study analyzes the implementation of XAI techniques, specifically the Local Interpretable Model-agnostic Explanations (LIME) algorithm (Ribeiro, Singh, and Guestrin [111]). LIME interprets complex model decisions by approximating the model locally with an interpretable linear model. By applying LIME, the aim is to make the decision-making processes of these models more transparent, thereby enhancing their utility for policymakers, while also providing human-understandable explanations of specific model outcomes (explanations deemed logical by the modeller). The explanations are judged on their human understandability, improving the explainability of the complex model, and validating the quality of the LIME explanations with objective metrics such as

the Coefficient of Variation and the Binary Classification Ratio.

This research focuses on the presence of street-level objects and/or conditions (cars, walkways, cleanliness, greenery, ..) in street views and how their presence contributes to the liveability of the public space. By analyzing street view images with XAI and incorporating the literature on liveability, assessments can be made about how possible objects and their relations influence urban liveability in street views.

Thus, the objective of this research is to generate human-understandable explanations of the decision behaviour of the CVDCM and to enhance the model's explainability. This objective contributes possibly to deeper knowledge about how humans value their street-level conditions, and what factors contribute to this. Policymakers can use these outcomes to make informed decisions about urban planning and public space management. Due to the cooperation between the TU Delft and the Municipality of Rotterdam, this model is applied to the city of Rotterdam to assess whether combining them with LIME can be useful for policy advice by mapping the city based on liveability through street view images. Additionally, the results are explained with LIME, to provide human understanding explanations for each street view and increase the explainability of the model's decision behaviour. In this regard, this research aims to bridge the gap between advanced Computer vision-enriched discrete choice models and practical, actionable insights for urban liveability via the LIME XAI algorithm for images.

A wicked problem is a complex issue that is difficult to define and inherently resistant to resolution due to its interconnected and multifaceted nature, involving numerous stakeholders and variables (Rittel and Webber [114]). The core wicked problem and engineering and policy relevance in this research is the opacity of AI models used for policy advice. This problem is wicked for several reasons. First, the inherent complexity of AI models makes their decision-making processes opaque and difficult to interpret. This opacity presents a significant challenge for policymakers who need to understand and justify the decisions made by these models to stakeholders. Second, the multifaceted nature of urban liveability compounds this issue. Liveability involves a wide range of factors, including safety, greenery, economic conditions, social cohesion, and accessibility to services (Khorrami et al. [63]), each of which can affect and be affected by other factors. These factors are not easily captured in street views alone, making comprehensive analysis challenging. Third, the implications of AI-based decisions in urban planning are significant and far-reaching, impacting all citizens, the environment, and the economy. The high stakes involved necessitate transparent and explainable AI models to ensure trust and acceptance among all stakeholders.

This research is structured as follows: Chapter 2 discusses the literature on what preferred and perceived urban liveability is, how computer vision models are used in urban research and how eXplainable Artificial Intelligence can help urban research provide explainability to the results from these urban computer vision models, Chapter 3 will discuss the methodology for the computer vision-enriched discrete choice model and Chapter 4 discusses the further application of LIME. Chapter 5 discusses the system and actor analyses of how such models can be used on the municipality level, Chapter 6 provides a face validity analysis of the CVDCM with an application to Rotterdam, with additional data validation, and in Chapter 7 the results of the eXplainable AI application is described for 7 case studies. In Chapter 8 and 9, this research is discussed and concluded.

## 1.1. Research question

To conduct this research, a central research question is formulated and several sub-questions are outlined to collectively address the main question. The research question is as follows:

- What benefits does the application of LIME towards computer vision enriched discrete choice models provide, and how can LIME explanations enhance their utility for policymakers in making informed urban planning decisions about the perception of the public space?

The scope of this research is on the application of the explainable AI technique LIME to the computer vision-enriched discrete choice models in street view images.

## 1.2. Sub questions

To answer the main question, several sub-questions need to be addressed:

1. **How is urban liveability defined and measured?**

Understanding liveability is vital for this research as it informs the possible influence on public space perception. The literature review will explore past analyses and existing measures, highlighting factors influencing urban liveability.

2. **How have Computer Vision models been used for modelling urban liveability and how is XAI used in combination with CV and liveability research?**

This question explores previous studies done on computer vision models that measure urban livability and defines the added value of using CV models for urban analysis. It also investigates the application of XAI to these models

3. **What face validity conclusions can we draw from the decision behaviour of the CVDCM without the application of XAI?**

The model is applied to selected neighbourhoods in Rotterdam. A face validity analysis examines possible behaviours in the model's outcomes, which are then analyzed, compared, and validated with municipal data.

4. **What challenges arise with the application of Street View Images for LIME image analysis?**

Using Street View Images with XAI computer vision models is new, and their application to algorithms like LIME has not been widely researched. This question identifies challenges in using street view images for LIME.

5. **How can the performance of LIME be measured?**

To provide scientific value to the quality of LIME explanations, it is important to identify metrics that can measure this quality.

6. **Can LIME enhance the explainability of the CVDCMs in comparison with the face validity study of the results?**

This question compares the results of both analyses and assesses whether LIME offers additional insights and explainability compared to the face validity study.

# 2

## Literature Review

Literature has been reviewed on human urban perception, computer vision models for urban analysis and using eXplainable Artificial Intelligence techniques in urban analysis.

### 2.1. Human Urban Perception

Within the context of urban analysis research, human urban perception refers to the subjective understanding, interpretation, and experience of urban environments by individuals or communities (Lynch [84]). It encompasses how people perceive various aspects of the city, including the physical characteristics, social dynamics, cultural attributes, and overall ambience (Caves [15]). In this literature review about human perception, first human perception itself is discussed, after which preferred liveability and perceived liveability are discussed, as well as the difference between the two and their importance in this research.

Human perception plays a crucial role in shaping people's interactions with their urban surroundings, influencing their attitudes, behaviours, and preferences. Researchers in urban analysis study perception to gain insights into how residents, visitors, and stakeholders perceive and interpret urban spaces, which can inform urban planning, design, and policy-making processes. Human perception research often employs qualitative and quantitative methods, including surveys, interviews, observations, and visual analysis, to capture and analyze people's perceptions and experiences (Porzi et al. [108]). By understanding human perception, researchers can identify strengths, weaknesses, opportunities, and challenges within urban environments, ultimately contributing to more informed and inclusive urban planning and development strategies. Human perception research in urban analytics can be broad, for example (Leby and Hashim [67]):

- **Physical Environment**  
How people perceive the physical features of the city, such as streetscapes, buildings, parks, and landmarks. This includes factors like aesthetics, scale, density, and architectural style.
- **Safety and Security**  
Perception of safety and security within urban areas, including concerns about crime, lighting, public spaces, and traffic.
- **Accessibility and Mobility**  
How people perceive the ease of movement within the city, including transportation options, pedestrian infrastructure, and the availability of amenities.
- **Social Environment**  
Perception of the social dynamics and interactions within urban communities, including diversity, inclusivity, social cohesion, and community engagement.
- **Quality of Life**  
Perception of overall well-being, satisfaction, and quality of life within urban environments, con-

sidering factors such as housing, healthcare, education, employment, and recreation.

Human perception is inherently subjective and varies among individuals and communities based on factors such as age, gender, socioeconomic status, cultural background, and personal experiences. It is therefore important to understand these differences in perception based on these factors. Human perception is dynamic and can change over time and across different spatial scales. For example, perceptions of a neighbourhood may evolve as it undergoes gentrification or revitalization efforts (Low and Altman [80]). Similarly, perceptions of safety and comfort may differ between daytime and nighttime or vary across different parts of the city. Human perception is influenced by psychological and emotional factors, such as cognitive biases, emotional responses, and sensory experiences. For instance, individuals may perceive certain urban spaces as welcoming or intimidating based on factors like architecture, street lighting, or the presence of greenery. There is a shared relationship between human perception and behaviour, where people's perceptions of the environment influence their actions and vice versa (Tuan [129]). For example, a perceived lack of safety may deter people from using public spaces, leading to underutilization and social isolation. Human perception research often draws on insights from various disciplines, including urban planning, psychology, sociology, geography, environmental psychology, and anthropology. This multidisciplinary approach allows researchers to explore the complex interplay between physical, social, cultural, and psychological factors that shape human perception. Findings from human perception research can inform urban design and planning practices by highlighting areas for improvement, guiding decision-making processes, and improving community engagement. For example, participatory design approaches involve collaborating with residents to co-create urban spaces that reflect their needs, values, and aspirations (Tuan [130]).

Human perception of place, as stated in Ordonez and Berg [105], captures the psychological connection residents have towards a neighbourhood. Cities, with their distinct functions, capture the emotional and psychological bonds urban buildings form with their habitat, as highlighted in Goodchild [43] and Dubey et al. [29], while also stating that the differences in built environments between places change also someone's sense of urban environment, thereby leading to diverse levels of psychological perception of citizens (Zhang et al. [145]). Such perceptions also benefit the semantic understanding of urban environments (Zhang et al. [145]). This underscores the importance of capturing and understanding these perceptions for urban planners and designers (Yao et al. [142]).

Liveability is closely connected to human perception as it reflects the subjective evaluation of various factors that contribute to the overall quality of life in urban environments. Human perception influences how citizens perceive the liveability of a city and the different neighbourhoods. Liveability encompasses various factors that contribute to the overall quality of life in urban environments, and human perception plays a crucial role in shaping how these factors are evaluated and experienced by residents and visitors. By understanding residents' perceptions of their urban environment, policymakers and urban planners can identify areas for improvement and develop strategies to enhance the liveability of cities and create more enjoyable and sustainable urban spaces.

Liveability is the concept used for determining the living qualities in cities. Cities are often divided into neighbourhoods by municipalities to facilitate organized distribution of goods, services, and effective city management, including policy interventions tailored to specific neighbourhoods (Myers [98]). This localized approach is deemed advantageous, marking the importance of focusing on neighbourhood quality of life rather than making cross-country comparisons (Shafer, Lee, and Turner [120]). Liveability, defined as a subjective evaluation by citizens of their living environment, reflects real experiences and contributes to a city's development, wealth, and prosperity (Shafer, Lee, and Turner [120]). Additionally, neighbourhood living standards significantly impact residents' health, with individuals in less favourable neighbourhoods facing elevated health risks and sub-optimal well-being (Barber et al. [8], Haan, Kaplan, and Camacho [47], Thompson and Kent [127]). The connection between neighbourhood well-being and health underscores the importance of monitoring and understanding these dynamics through a liveability framework (Evans [32]). Leidelmeijer and Kamp [70] defines liveability as the degree to which the environment aligns with the requirements and desires expressed by humans. The definitions and factors discussed in the literature that influence liveability are mainly factors that are 'perceived'. Therefore, most liveability research is about perceived liveability. This research will focus not on perceived liveability, but on preferred liveability.

### 2.1.1. Preferred liveability

Preferred liveability focuses on individuals' stated preferences or desires regarding the characteristics of their ideal living environment. It involves understanding what people value and prioritize in their living environment, regardless of whether those preferences align with their current circumstances or the actual conditions of where they live. Preferred liveability assessments often involve asking individuals to rate different features or attributes of a neighbourhood according to their importance and desirability. For example, citizens could be asked to choose one street view image between two images for which they would be most likely to relocate, then if a model is trained on these choices, a preferred liveability is analyzed of the citizens.

Preferred liveability drives behaviour by influencing individuals' decisions about where to live based on their aspirations and desired living conditions. When people identify a neighbourhood that aligns more closely with their preferences for factors such as safety, aesthetics, amenities, and community, they are more likely to relocate to that area. This decision-making process is guided by the theory of choice behaviour, which examines how individuals make trade-offs to achieve their desired outcomes (Samuelson [117], Luce, Jessop, and Berkeley [81]). When people consider moving, they weigh their current living situation against their ideal preferences. If a substantial gap exists between perceived liveability (their current conditions) and preferred liveability (their ideal conditions), it can motivate behaviour change, such as relocating to a more desirable neighbourhood. This relocation is driven by the need to improve life satisfaction and overall well-being, which are closely linked to living in an environment that meets their preferred criteria (Clark and Ledwith [18]).

This preferred liveability is defined as a residential location choice in this research. It portrays how preferable a certain street view is for citizens to relocate there. This preferred liveability, the residential location choice, is built on multiple factors described in the human perception research. Ewing and Handy [34] showed that measuring neighbourhoods solely on objective and physical features alone is not able to model the perception fully, as human perception of street quality contains subtle relationships that can't be measured solely by individual objects. This is also supported by Xu et al. [141], who state that only objective features, as analyzed with perceived liveability, cannot fully model the human perception of the street view. It encompasses complex relations between the factors of what liveability is. Citizens' preferences are studied in the theory of choice behaviour (Samuelson [117], Luce, Jessop, and Berkeley [81]) and determine individuals' selections and decision-making processes involving trade-offs, such as residential location choice. In this scene, perceptions are subjective interpretations of factors in the neighbourhood. While they may impact individuals' choices, perceptions do not inherently dictate them. Therefore, perceived liveability helps us understand what could influence the preferred liveability.

Perceived liveability refers to individuals' subjective perceptions or evaluations of the quality of life and overall liveability of a particular place or environment. It is based on people's personal experiences, observations, and interpretations of various factors that contribute to their satisfaction with their living conditions. Perceived liveability is often measured through surveys, interviews, or self-reporting methods to capture individuals' subjective assessments of their living environment. Perceived liveability is more static and represents an individual's assessment of their current living environment. It is based on personal experiences and observations of factors such as safety, noise levels, access to services, and community engagement. While perceived liveability impacts satisfaction and quality of life, it does not inherently drive behaviour because it is an assessment of the status quo.

In contrast, preferred liveability is dynamic and represents the ideal or aspirational living conditions individuals seek. This aspirational nature drives behaviour as people are motivated to bridge the gap between their current living conditions and their desired ones. Therefore, preferred liveability is a proactive driver of change, influencing decisions like moving to a new location, whereas perceived liveability is a reactive measure, reflecting satisfaction with the current situation (Ewing and Handy [34]).

Understanding the distinction between perceived and preferred liveability is crucial for urban planners and policymakers. By recognizing what people aspire to in their living environments, planners can design and develop urban spaces that align more closely with residents' preferences, thereby improving

satisfaction and reducing the need for relocation. This understanding can help create more sustainable and cohesive communities where people are more likely to stay and invest in their neighbourhoods. Additionally, recognizing that preferred liveability drives behaviour allows policymakers to address disparities between different areas. If certain neighbourhoods consistently fall short of residents' preferred liveability criteria, targeted interventions can be implemented to enhance those areas, thereby balancing development and improving overall urban liveability (Van Herzele and Wiedemann [133]).

By focusing on preferred liveability, urban development can become more attuned to residents' needs and desires, leading to healthier, happier, and more stable communities. This approach not only enhances individual well-being but also contributes to the broader goals of urban sustainability and social equity.

### 2.1.2. Perceived liveability

Perceived liveability refers to individuals' subjective perceptions or evaluations of the quality of life and overall liveability of a particular place or environment. It is based on people's personal experiences, observations, and interpretations of various factors that contribute to their satisfaction with their living conditions. Perceived liveability is often measured through surveys, interviews, or self-reporting methods to capture individuals' subjective assessments of their living environment.

Reviewing perceived liveability alongside preferred liveability is integral because it provides an empirical basis for understanding the factors that contribute to a desirable living environment, thereby influencing the preferred liveability. After model usage, findings can be grounded in the literature review on perceived liveability, identifying factors also found in the literature, and ensuring that the model's assessments align with the aspects of urban environments that residents value most.

A thorough literature review on perceived liveability allows for identifying key features that impact residents' satisfaction with their living conditions. Insights from perceived liveability studies assist in the later analysis of the model's outcomes, as these can be related to possible features identified in the literature. All features listed in this literature review answer the first subquestion: 'How is liveability in neighbourhoods defined and with which indicators is it measured?'. The literature review on perceived liveability is not just about understanding current satisfaction but also about using that understanding to interpret and validate the results of the preferred liveability model effectively.

The review focuses on humanly detectable features in street views. This focus is important because the analysis in this research will be on visual elements that can be identified and interpreted by both human and computer vision models. Elements like openness, enclosure, continuity, and cross-sectional proportion, while theoretically significant (Owens [106], Ewing and Handy [33], Alexander et al. [3], Dover and Massengale [28], Montgomery [95], Harvey [50], Heath et al. [53], Gehl and Svarre [40], Harvey et al. [52], Gehl [39], Bruce, Green, and Georgeson [11], Li, Ratti, and Seiferling [73]) are abstract concepts that are not easily quantifiable or visually detectable in street views by humans. By concentrating on tangible features that are readily observable in street views, the model's outputs can be meaningfully related to real-world objects or clusters of objects.

The research by Tang and Long [126] defines the attributes of the visual street quality, which are taken from Harvey et al. [52] and Saelens, Sallis, and Frank [116], emphasizing the importance of greenery. Greenery is defined as the amount of green present in the streetscape (trees, bushes, and plants for example), which has a positive effect (increasing the amount of greenery) on the feeling of people (Li, Zhang, and Li [74], Li et al. [75]). Kirdar and Cagdas [64] defines image-liveability as related to the visual quality of streetscapes, with measurement variables including urban image evaluation (liveability), urban greening, building conditions, physical incivilities affecting street legibility, traffic density, and building enclosure rate. Urban image evaluation is crucial, as studies such as Nasar [101] have demonstrated that the visual quality of cities significantly impacts people's feelings in a city. Urban greening focuses on the positive effects of nearby green spaces on liveability. Building conditions are important, as individuals perceive clean and well-maintained buildings as contributing positively to liveability. The presence of physical incivilities that deface street legibility, such as undesirable traffic signs, negatively influences people's perceptions of their street view. Traffic density, as indicated by Appleyard, Gerson,



and Lintell [5], also affects how people perceive the attractiveness of their streetscape.

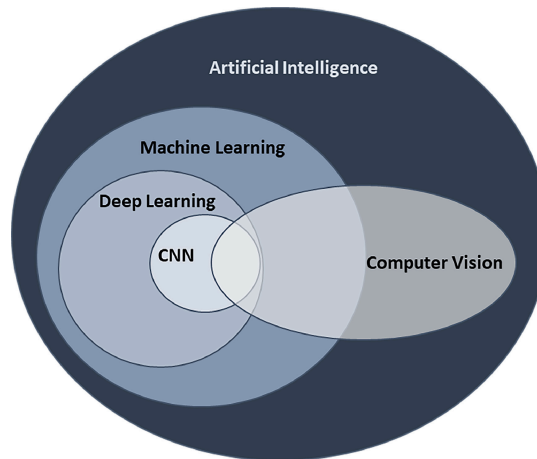
Harvey and Aultman-Hall [51] emphasize the importance of the non-visual effects of street views on liveability, which are influenced by many factors, such as human activity and weather, and the amount of light there is in a streetscape (Nasar [102], Cassidy [14]). Nasar [103] analyzed the human perception of the environment in cities and found that greenery, order, openness, historical significance, and upkeep were the most important features for improving the human perception of the neighbourhood. Weber, Schnier, and Jacobsen [138] found that consistent architectural style, symmetry, scale, and greenery are major influences on the human perception of public spaces. Kelling and Coles [62] and Skogan [122] showed that covered-up and/or abandoned buildings and cars, trash, and litter influence human perception negatively. Khorrami et al. [63] reviews the literature on perceived liveability factors and in addition to the research already discussed, finds that the height of homes, the number of open spaces, cleanliness, the road conditions, the ability to walk and bike, the number of commercial stores, the presence of greenery, water and trees, the quality of the pavements, the number of vehicles, and the amount of lighting influence human perception of the neighbourhood.

Walkability, the presence of pedestrian pathways, and accessibility to public transport significantly contribute to perceived liveability (Gehl [39], Jacobs [58]). Well-maintained roads and pavements, the presence of street furniture like benches, and adequate lighting also play crucial roles in enhancing perceived safety and comfort, contributing to overall liveability (Ewing et al. [35]). Moreover, the diversity of land use, such as a mix of residential, commercial, and recreational spaces, positively impacts perceived liveability by providing convenience and fostering community interaction (Montgomery [96], Handy et al. [49]). The aesthetic quality of buildings, including architectural design and the upkeep of facades, has been found to influence residents' satisfaction with their environment (Nasar [103], Talen and Koschinsky [125]).

Worldwide, there are three most used perceived liveability measures to indicate liveability worldwide, and in the Netherlands, the Leefbarometer is a national tool to analyze liveability nationwide. All four liveability frameworks are further discussed in Appendix A, which are provided as additional interesting information for the reader. Human urban perception is defined by how people perceive the urban space, while preferred liveability allows for analyzing the factors that play a role in behaviour change in location change, and perceived liveability provides insights into the factors affecting the perceived liveability. As can be seen, a clear uniform definition is not present, while the distinction between preferred and perceived liveability is easily made, and the features on which to measure it are in abundance, no clear set exists.

## 2.2. Computer vision in Urban research

After conducting a comprehensive literature review on perceived and preferred liveability, it is crucial to review the application of computer vision models in urban research, as these type of models provide an innovative technique for modelling urban liveability. Computer vision is a field of artificial intelligence that enables computers to interpret and understand visual information from the real world, such as images and videos. It involves developing algorithms and techniques to analyze, process, and extract meaningful insights from visual data, allowing machines to perceive their environment similarly to humans. Computer vision finds applications in various domains, including object recognition, image classification, facial recognition, medical imaging, autonomous vehicles, and surveillance systems (Marasinghe et al. [87]). Wiley and Lucas [139] define computer vision as the technology that allows a computer to analyze visual data and extract relevant information, which can help us understand the urban environment. Figure 2.1 shows the artificial intelligence field in which computer vision acts.



**Figure 2.1:** The position of Computer Vision in the Artificial Intelligence field (Marasinghe et al. [87]).

The research of Yigitcanlar et al. [143] showed that the use of AI technologies in urban planning is not yet vastly used in comparison with other sectors such as medication and finance. Incorporating AI into urban research methodologies offers urban planners a chance to analyze vast urban datasets and identify patterns and trends via advanced modelling and simulation techniques which could be beneficial for urban planning (Son et al. [123]). Computer vision models can analyze visual data, which is important for determining information in urban processes (Liang, Zhao, and Biljecki [77] Guzder-Williams et al. [46]). Yin and Wang [144], for example, was one of the first to research the application of computer vision and machine learning in urban analysis, for obtaining street features from street view images and measuring the visual enclosure of streets. Naik [99] and Shi [121] showed the potential and capabilities of using CV in urban planning by using it for extracting information from images of landscapes.

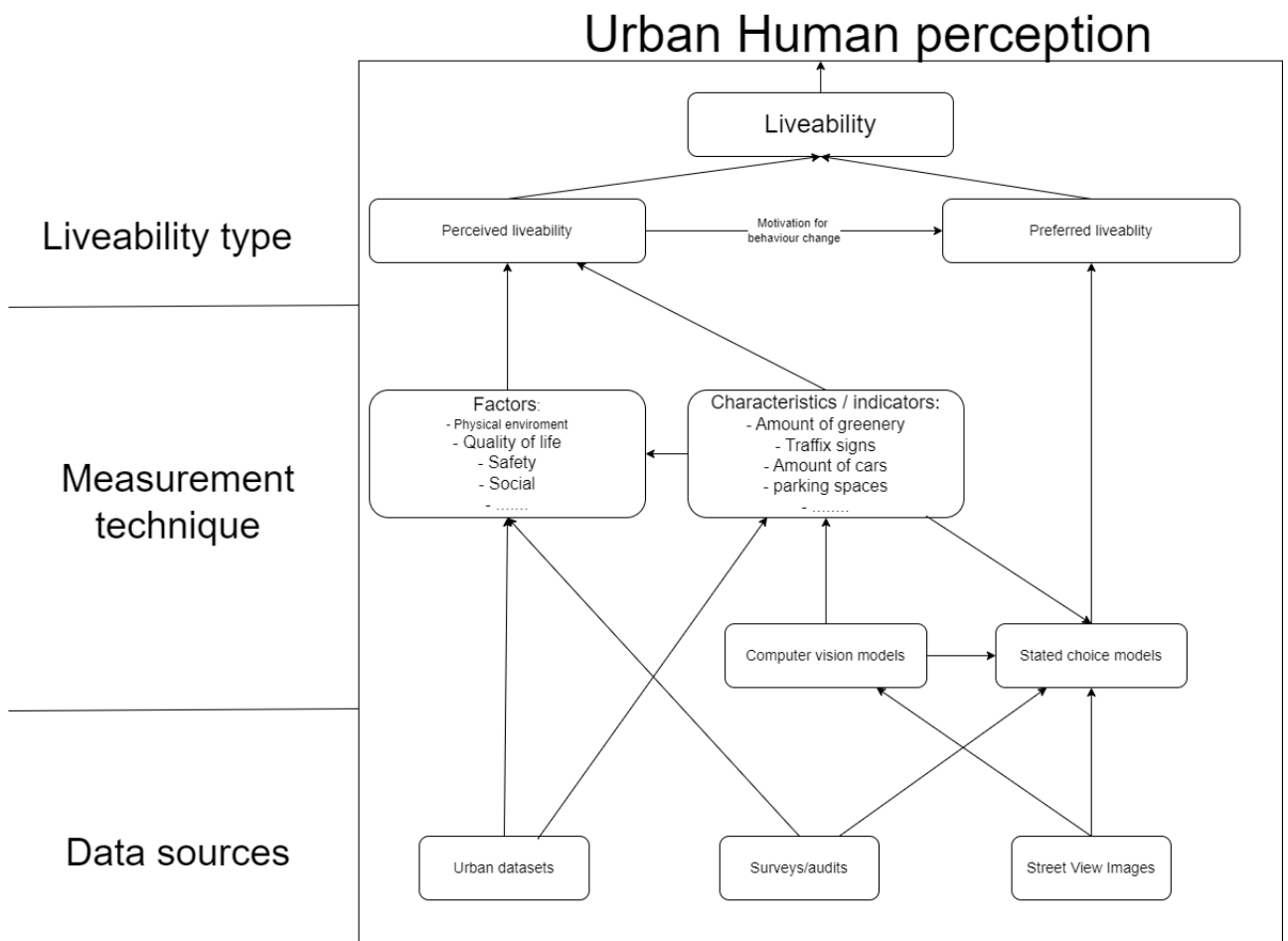
In urban research with computer vision research, street view images are vastly used as input data for computer vision models. Badland et al. [7] showed that the use of street view images allows for not doing field visits anymore, which are costly, time-consuming, and require manpower, while the use of street view images allows for virtual audits, is cost-effective, require no physical people examining the streets, and is vastly present, therefore not time-consuming. Field surveys were previously the method for collecting urban data, which was time-consuming, costly, and required manpower. Street view images are from the eye level of citizens, which allows for better human perception research, as visual perception can be captured with street view images. Multiple researches have been done on the amount of green in cities and the effect it has on human perception, safety and other factors. Transportation and mobility is another research area where street view images are used. Since most street view images are taken from the road, and also view the road, they can be used for analyzing mobility networks and transportation opportunities (Najafizadeh and Froehlich [100]). Next to transportation and mobility, walkability is another concept researched in combination with street-view images (Ito and Biljecki [56]), where analyzing street views can interpret the walkability of streets and cities by analyzing the room available for walking and the quality of these streetwalks. Zhang et al. [145] used computer vision for object detection in street views and analyzed the presence of certain objects in the human perception, allowing to investigate the possible influence of certain street objects on the human perception of public space. In Li, Zhang, and Li [74], street view images are used for analyzing the relation between perceived safety in street views and the amount of greenery present, using object segmentation to subtract the amount of greenery from each street view image. Chen [16] investigates the perception of public open spaces in Hong Kong and Singapore. They use a stated choice experiment, where participants need to choose between two street views for which one provides the most utility on 6 different levels: accessibility, amenities and furniture, design and aesthetics, environment, safety, and use and user. With these responses, liveability concerning street views is analyzed.

As can be seen, the possibilities of using computer vision for urban analysis are vast, and much research has been done in recent years, as it allows new insights into urban planning (Marasinghe et al. [87]). Computer vision allows modellers to subtract relevant information from street views and relate

this to liveability, or allow the use of stated choice experiments for modelling preferred liveability in street views.

### 2.2.1. Human perception overview

Figure 2.2 provides a visual summary of the relationship between perceived and preferred liveability and the methods used to measure them. It illustrates that three primary data sources contribute to liveability research: urban datasets, surveys, and street-view images. Urban datasets offer information on various urban factors such as infrastructure, green spaces, and socio-economic conditions, helping to understand the physical and social characteristics of urban areas. Surveys gather subjective data directly from individuals regarding their experiences and preferences, which is crucial for capturing perceived liveability and understanding the factors that influence it. Street-view images, analyzed through computer vision techniques, provide insights into the urban environment by detecting and interpreting features within the images. This analysis can include object detection, spatial arrangement, and overall image assessment, thereby contributing to both perceived and preferred liveability metrics. The figure also highlights two key measurement techniques: computer vision and stated choice models. Computer vision is used to analyze street-view images, facilitating the detection of objects and overall assessment of urban conditions. Stated choice models, on the other hand, are employed to model the preferred liveability choices of individuals, thus directly analyzing their ideal urban conditions. Furthermore, the distinction between direct characteristics and factors is important in this context. Direct characteristics, which come from urban datasets or computer vision models, are quantitative attributes of the urban environment. Factors, which are typically captured through surveys and urban datasets, represent qualitative aspects that serve as proxies for perceived liveability. Understanding the difference between perceived and preferred liveability is crucial as it sheds light on the motivations behind potential behavioural changes, such as relocating homes, aimed at achieving a more desirable urban environment. Both perceived and preferred liveability are essential for gaining a comprehensive understanding and for enhancing urban liveability.



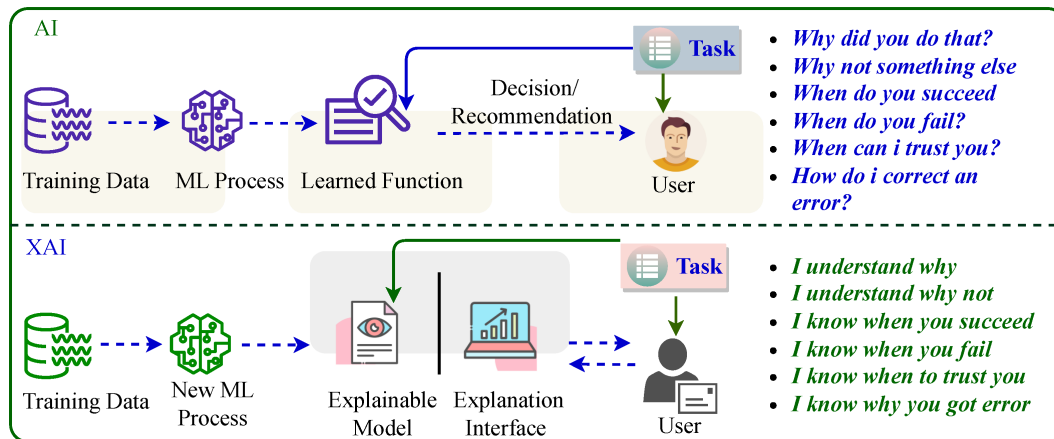
**Figure 2.2:** Human Perception overview showing perceived and preferred liveability, how they are measured and from which data they originate

### 2.3. eXplainable AI in urban research

The integration of an eXplainable AI (XAI) literature review following the computer vision literature review is critical as it allows for understanding the decision behaviour of the computer vision models, which improves possible decision-making for policy advice with these models. Artificial Intelligence algorithms are generally composed of complex nonlinear relationships. Attempting to examine and articulate the decision behaviour of these algorithms is practically impossible, and results in not fully understanding an AI's decision-making process. As a result, AI algorithms are frequently characterised as a 'black box', as determining the decision behaviour of the model is not possible and thus lacks explainability. The benefit of these artificial intelligence models is that they can model complex relations we as humans are not able to model ourselves. These AI algorithms (deep neural networks) are highly efficient predictors, which come from the complex non-linear statistical models and the number of parameters present in these models. This leads to low transparency of the model's behaviour (Li et al. [72]). This results in opacity, which means that the model is unable to provide reason and explanation for the decisions it has made. A black-box model is a machine-learning model functioning as an opaque system. In such models, the internal mechanisms are not readily accessible or interpretable. While these models generate predictions using input data, the user is unable to discern the transparent decision-making process or the rationale behind the predictions (Garg et al. [37]).

To deal with this opacity, XAI techniques are used to provide explanations. XAI is a recent study field that has gained a lot of attention and further research (Gunning and Aha [45]). Arrieta et al. [6] defines XAI as a system that produces details or reasons to make a models functioning clear or easy to understand, while Das and Rad [24] defines XAI as a field of AI that promotes a set of tools, techniques, and

algorithms that can generate high-quality interpretable, intuitive, human-understandable explanations of AI decisions, and Adadi and Berrada [2] refers to XAI as the movement, initiatives, and efforts made in response to AI transparency and trust concerns, more than to a formal technical concept. As can be seen, the definitions represent the same understanding, while the explanations are different. They characterize XAI as a collection of machine learning methods that empower human users to comprehend, confidently trust, and proficiently oversee AI models. Keywords for XAI are interpretability and explainability, which are interchangeably used in the literature. Miller [93] defines interpretability as “the degree to which an observer can understand the cause of a decision” and states that the definitions of interpretability and explainability are the same. In the case of using AI algorithms for policy advice, explainability can increase the trust for made policy decisions (Ehsan et al. [31]). Figure 2.3 shows the difference between the working of ‘normal’ AI models and XAI models.



**Figure 2.3:** The difference between the usage of ‘original’ AI models and XAI models (Javed et al. [59]).

In the ‘normal’ AI model, the user analyses the results without knowing how the model came to the decision, leaving the user with questions such as: ‘How did the model do that?’, and ‘Why not something else?’. In the second case, the user interacts with the model to understand why a certain outcome has been created, enabling him to understand why this certain street view image gained a certain outcome and creating trust in the decision process of the model.

The analysis of the explanation of computer vision urban research has been mostly focussed on feature importance analysis of models’ objective features. In some cases, specific XAI techniques have been used, such as in ujatha, Lavanya, and Prakash [131], where they use Shapley values as an XAI technique to analyze the decision behaviour of a street view-based computer vision model with predefined objects to measure the human perception. Lee, Kim, and Park [69] analyses the urban environment for walkability with computer vision, where again Shapley values are applied as XAI technique, on features predefined in the model. Using prior object detection is exemplary of all human perception/liveability research with street view images found in the literature.

Law et al. [66] discusses eXplainable AI techniques for computer vision models in urban research. It explains that XAI techniques can pose challenges in highlighting the factors upon which these models base their decision behaviour. The generated explanation can contain multiple objects in the image or only certain parts of objects when applying it to street view images. This problem of multiple interpretation possibilities is called holistic image tasks. In the human perception of the urban environment, objects/factors of the street views are highly related to each other, as multiple factors influence the liveability and are present closely to each other (trees, cars, buildings, sky). They state that with the increasing use of computer vision models for urban research, it is expected that more scientists/ urban analysts (because urban analytics are most of the time not computer vision experts) will use them, without deep knowledge about their workings. Hence, it is crucial to offer explainability tools enabling these scientists to develop insights into system operations, mitigating the risk of misinterpretation of

their outputs and creating un-well supported policies. Kakogeorgiou and Karantzalos [60] states that the advantage of using prior object segmentation in urban research with computer vision is that it is much easier to measure and interpret the results, as it is inherently objective. The measurements are based on predefined objects, and their ratio is assumed to influence the research variable. Therefore, the results are clearer and explain factors already commonly known and thus interpretable. This can explain the lack of applied XAI techniques in the field of urban research with computer vision. In this part of the literature review about computer vision and XAI for urban research, the second subquestion is answered: 'How have Computer Vision models been used for modelling urban liveability and how is XAI used in combination with CV and liveability research'. In a vast array of research topics has computer vision been used, especially in combination with street-view images. XAI on the other hand has not been used much, where mainly predefined feature importance is taken for explainability of the results.

The quality of an XAI explanation is critical because it directly impacts the user's ability to understand and trust the AI model's decisions. Doshi-Velez and Kim [27] states that an explanation should be intelligible and tailored to the audience's expertise level, ensuring that the insights derived from the model are accessible and actionable. This means that the explanations need to be simple enough for non-experts to grasp yet detailed enough to provide meaningful insights to experts. Furthermore, the quality of explanations must be evaluated to ensure they are genuinely informative and not misleading. Lipton [79] emphasizes that explanations should be faithful to the model's processes, meaning they accurately represent the model's reasoning and are not just plausible narratives that humans find easy to understand. This fidelity is crucial for maintaining the integrity of the decision-making process and ensuring that the explanations are reliable. The necessity of high-quality explanations is underscored by the potential consequences of poor-quality or misleading explanations. If stakeholders cannot trust the explanations, the overall trust in the AI system diminishes, potentially leading to the rejection of valuable insights or, worse, the implementation of flawed decisions based on misunderstood data (Ribeiro, Singh, and Guestrin [111], Lundberg and Lee [82]). Moreover, Miller [93] discusses the importance of coherence in explanations, where the explanation should logically follow from the data and the model's operations. Coherent explanations help users build a mental model of how the AI system works, which is essential for trust and usability. This aspect is particularly important in complex domains like urban liveability, where multiple factors interact intricately.

Incorporating high-quality XAI explanations into research enhances the overall interpretability of AI models. This dual focus ensures that the models are transparent and their outputs are meaningfully understood and effectively used. Therefore, besides enhancing trust and interpretability, the quality of XAI explanations themselves must be rigorously evaluated. This ensures that the explanations are not only accurate and faithful to the model but also understandable and useful to the intended audience. This comprehensive approach to XAI helps bridge the gap between advanced AI capabilities and their practical, real-world applications, ultimately leading to better-informed decisions and more effective policies.

## 2.4. Research gap: XAI for urban research with computer vision models

In most urban research utilizing computer vision, prior object detection is used (Marasinghe et al. [87]). These methods rely on predefined objects within the segmentation model, and the outcomes are based on these predefined objects. This approach is effective when the research topic can be objectively measured, such as the quantification of greenery, which can be clearly defined by the presence of trees, bushes, and plants (Li and Ma [76]). This method allows for the model's decision behaviour to be clearly explained. However, for more abstract concepts like preferred liveability, there are no universally agreed-upon objective measures. While perceived liveability literature suggests potential objective measures, there is no definitive set for preferred liveability. Computer vision models, being opaque systems, can benefit from object segmentation to explain their outcomes. However, relying on predefined objects introduces certain assumptions and biases in the research, as the conclusions are drawn from these predefined measures (Gong et al. [41], Doiron et al. [26], Chen and Biljecki [17]). This approach limits the scope of understanding to the predefined objects and does not fully capture

the holistic human perception of liveability, which involves complex connections beyond individual view indices or their combinations. Human perceptions are intricately connected to visual elements in ways that cannot be entirely captured by predefined objects alone (Ewing and Handy [34], Lin and Moudon [78]).

This research addresses the gap by not predefining which objects will be measured, but instead applying XAI techniques to understand the model's outputs after processing the street view images. This approach maintains the holistic view of human perception and allows for the analysis of complex relationships between all objects in street views (Qiu et al. [109]). XAI helps in explaining why certain liveability scores have been generated, providing a more nuanced understanding of what influences liveability without the bias of predefined objects. Law et al. [66] emphasize that creating a human-understandable explanation in complex scenarios, such as street view images, where there are multiple important objects in the image, is hard. Since these explanations have to increase the interpretability of the complex model, but also need to be humanly understandable, the complexity of street view images is tackled with explanations focussing on the objects in street views, allowing it to be related to the literature on liveability. This thus eliminates the prior need for object detection, while using object detection qualities post modelling for explanations. These post-explanation techniques allow for more interpretable visual information, which can be valuable for policymakers. By using XAI, this research aims to enhance the explainability of computer vision models in urban analysis, potentially providing new insights into human perception of street views by analyzing the holistic human perception of street views' liveability. The visual explanations from XAI can serve as effective tools for policy engagement, aiding in the formulation of informed urban policies.

# 3

## Methodology

This chapter presents the methodology for the Computer Vision-enriched Discrete Choice Model from Cranenburgh and Garrido-Valenzuela [23], aimed at evaluating urban liveability using street view images, and the application of LIME, the XAI technique. The insights gained from the literature review informed the methodological choices outlined in this chapter, with an understanding of the limitations and advantages of existing approaches guiding the selection of appropriate methods for our research. The research design adopted for this study is quantitative, combining descriptive and exploratory approaches to gain a comprehensive understanding of urban liveability. The methodology encompasses the theoretical framework, data collection processes, and analytical techniques.

### 3.1. Data Retrieval for model usage

Street-view images must be collected to acquire the requisite input data for model usage. This process involves overlaying a grid onto the geographical map of the research area, whether it is a city, town, or an entire country. The spacing between these grid lines is determined by the researcher, with a set distance of 50 meters chosen for this study. As a result, this creates intersection points within the grid, where the horizontal and vertical lines of the grid intersect. For all these intersection points falling within the research area, street-view images are gathered. Each intersection point is associated with the same chosen radius, representing a geographic area. Within this circular area around the intersection point, all available street-view images are retrieved. For this study, a radius of 200 meters is employed. The research area for this thesis is certain areas in Rotterdam which yielded over 300,000 images, necessitating significant data processing. Thus, the chosen radius and grid line distance, not only influence dataset granularity but also impact the volume of data to be collected.

At each location where a Street View image is captured, four images are taken: left, right, front, and back. This creates a panoramic view of the location. Consequently, these photos thus offer views of the house's surroundings from various angles. Specifically, they provide perspectives of the house's front, the view from the house's front, and the streets to the left and right of the house. This study operates under the assumption that the perception of outdoor space is captured by the street view showcasing the house and the surroundings. Given that street view photos are captured for every point within the specified radius, the remaining street views depict the surroundings of neighbouring properties. Notably, these views include the front of neighbouring houses and the corresponding perspectives, enriching the dataset with contextual information.

### 3.2. Computer-vision enriched discrete choice model

The model described in Cranenburgh and Garrido-Valenzuela [23] is a Computer Vision-enriched Discrete Choice Model (CV-DCM), which is a combination of a computer vision model with a traditional discrete choice model involving both numeric attributes and visual imagery. A discrete choice model is a statistical model used to analyze and predict decisions made by individuals when presented with a set of options. Here, the preferred liveability as discussed in the literature review comes into play.



It quantifies how people choose among alternatives based on the attributes of those options and the preferences of the decision-maker. This modelling approach is used to analyze human behaviour in choosing between discrete alternatives. Computer Vision-enriched Discrete Choice Models integrate principles of Random Utility Maximization (RUM) with computer vision techniques to analyze decision-making processes. Unlike traditional Discrete Choice Models (DCMs), which are limited to numeric data, CV-DCMs can handle choice tasks that involve images (street view), allowing analysis of not only the decision behaviour based on numerical values but also on street view images. Therefore, these types of models suit well for modelling preferred liveability in combination with street view images.

Random Utility Maximization is a foundational concept in the field of choice modelling and serves as the theoretical basis for Discrete Choice Models (DCMs) (Manski [86]). Utility refers to the subjective satisfaction or preference that an individual derives from choosing a particular alternative. It represents the perceived value or benefit associated with selecting one option over others (these discrete options). Random utility acknowledges that individuals make choices based on both observed attributes of alternatives and unobserved or random factors that influence their preferences. These random factors could include individual-specific characteristics, unobservable attributes of alternatives, or stochastic elements in decision-making. Random Utility Maximization states when individuals choose the alternative that maximizes their expected utility. In other words, when faced with multiple options, individuals will select the alternative that they believe will provide them with the highest level of satisfaction or utility, given the available information and their preferences. In mathematical terms, Random Utility Maximization is typically expressed through a choice probability model. These models quantify the probability that an individual will choose a particular alternative from a set of options based on the observed attributes and the individual-specific random component (Cascetta [13]).



In the research of Cranenburgh and Garrido-Valenzuela [23], the influence and effect of street views on the location choice behaviour of citizens are emphasized concerning travel time, cost and the street view of the location. A stated choice experiment was used to gather response data, which encompassed the street view images in combination with added travel time and costs (numeric values), and let participants choose between two options. This experiment can be seen in Figure 3.1. In this stated choice experiment, participants were instructed to envision moving to a different home. They were then presented with two residential options (street view images) and asked to select their preferred choice. Before making this choice, participants were provided with the following information:

- The new house would be identical to their current one in terms of size, type, year built, furniture, maintenance, etc., with only the neighbourhood/public space changing.
- Monthly housing costs (including rent, mortgage, taxes, insurance, etc.) could either increase or decrease.
- The new neighbourhood would be relatively close to their current one, but commute time might change. Commute time was based on their current mode of transportation.
- All other aspects of their situation would remain the same, including distances to amenities, schools, healthcare facilities, etc.
- The images shown in the choice tasks depicted the ground-level view from the street side.

The survey choice experiment:

Suppose, you have to relocate to a different neighbourhood. Your house stays the same; only the neighbourhood changes. You have two options.

**Which option would you choose?**

	Option A	Option B
Your new street-view		
Monthly housing cost	€0 equally expensive as present	↑ €225 more expensive than presently
Commute travel time	↓ 5 minutes quicker than presently	↓ 10 minutes quicker than presently
	<input type="radio"/> Option A	<input type="radio"/> Option B

**Figure 3.1:** Screenshot of the participants stated choice experiment from Cranenburgh and Garrido-Valenzuela [23]

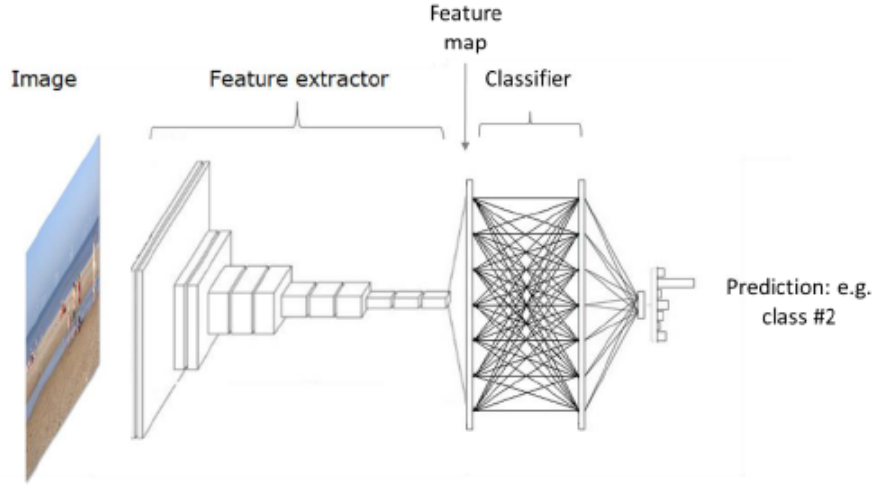
The survey was held in multiple municipalities in the Netherlands, providing a sufficient number of choice outcomes connected to street view images to train the model. This also results in that the model represents the preferred liveability of the average Dutch citizen. The survey has been filled in by a diverse set of people, but the model outcome is for all the preferences combined, therefore modelling the preferred liveability of the average Dutch citizens.

The stated choice experiment described above effectively models the preferred liveability as discussed in the literature review. By utilizing Random Utility Maximization, the experiment quantifies the ideal living conditions based on participants' choices between two street view images. This approach captures the essence of preferred liveability by analyzing how individuals weigh and prioritize street views in their decision-making process. However, it is important to note that this method thus does not model perceived liveability. Perceived liveability refers to the subjective evaluation of an environment based on an individual's current experiences and impressions. In contrast, the stated choice experiment focuses on the ideal conditions or preferences individuals express when comparing discrete options, rather than their ongoing perceptions of these conditions.

Next to the modelling of the choice behaviour, the training of the model also includes street view images. These images, comprised of raw pixel values, possess dimensions determined by their width, height, and colour depth, typically represented in Red, Green, and Blue (RGB). However, the sheer volume of pixels, often numbering in the millions, results in high-definition images with varying colours. For instance, a seemingly straightforward 100x100-pixel image with three colour channels already encompasses 30,000 pixels, showcasing the complexity even in seemingly modest images.

Manipulating images at the pixel level is neither efficient nor effective for computer analysis. This is because the high dimensionality of detailed images increases computational demands, spatial relationships within the image are lost, and robustness to variations such as changes in lighting and viewpoint is reduced when analyzing individual pixels. Therefore, a convolutional neural network (CNN) is used to process the street-view images. CNNs extract important features from raw pixel values through a series of layers, capturing spatial relationships and highlighting task-specific representations directly from the data. This approach not only reduces dimensionality issues but also improves the model's ability to handle different image variations (LeCun et al. [68]). The utility scores obtained from the stated

choice experiment, alongside the images themselves, serve as training data. By passing street view images through the CNN, the model extracts meaningful features from the images instead of having to work with individual pixel values. This feature map generally can provide a representation that can be mapped linearly (Goodfellow, Bengio, and Courville [44]). Figure 3.2 shows a common architecture of how a CNN is built.



**Figure 3.2:** The architecture of the Convolutional Neural Network of the CVDCM from the research of Cranenburgh and Garrido-Valenzuela [23]

The model is a machine learning model trained on the stated choice experiment results and the computer vision CNN model for subtracting the feature map of images. In this model, the linear layers allow further processing of these extracted features alongside additional attributes such as cost, and commute travel time, as seen in the stated choice experiment. This combination provides the computation of utility values associated with both the image features and these supplementary attributes.

The calculated score in this model will from now on be called the utility-based liveability score, as it is the score representing the liveability based on the utility gained from the stated choice experiment. To understand how this utility-based liveability score is calculated, a more technical view of this model is needed. Note that the following explanation is directly taken from Cranenburgh and Garrido-Valenzuela [23], for a more in-depth analysis of the model, it is suggested to first read his paper.

Consider a decision-maker,  $n$ , who faces a multi-attribute choice task with a set of  $J$  mutually exclusive alternatives. Each alternative is described by  $M$  numeric attributes  $X_j = \{x_{j1}, x_{j2}, \dots, x_{jM}\}$ , such as the travel cost and travel time and by a (colour) image  $S_j$  with a resolution of  $H \times W \times C$ .

We assume this decision-maker makes decisions based on this Random Utility Maximising (RUM) principle, which is stated in equation 3.1, where  $U_{jn}$  denotes the total utility experienced by person  $n$  considering alternative  $j$ ,  $V_{jn}$  is the utility experienced by decision-maker  $n$  derived from attributes observable by the analyst. And, to account for the fact that the analyst does not observe everything that matters to the decision-maker's utility, an additive error term  $\varepsilon_{jn}$  is added to each alternative (Train [128]).

$$U_{jn} = V_{jn} + \varepsilon_{jn} \quad (3.1)$$

Furthermore, it is assumed decision-makers experience utility from both the numeric attributes  $X_j$  and image  $S_j$ , see Equation 3.2, where  $v$  is a preference function which maps the attributes and image of an alternative onto the utility.

$$U_{jn}(X_{jn}, S_{jn}) = v(X_{jn}, S_{jn}) + \varepsilon_{jn} \quad (3.2)$$

Three assumptions are further made to calculate the overall utility. First, the utility derived from the numeric attributes and the image are separable and additive in the utility space. This is given in equation

3.3, where function  $f$  maps the (observed) numeric attributes onto the utility and function  $g$  maps the relevant information from the image onto the utility.

$$U_{jn}(X_{jn}, S_{jn}) = f(X_{jn}) + g(S_{jn}) + \varepsilon_{jn} \quad (3.3)$$

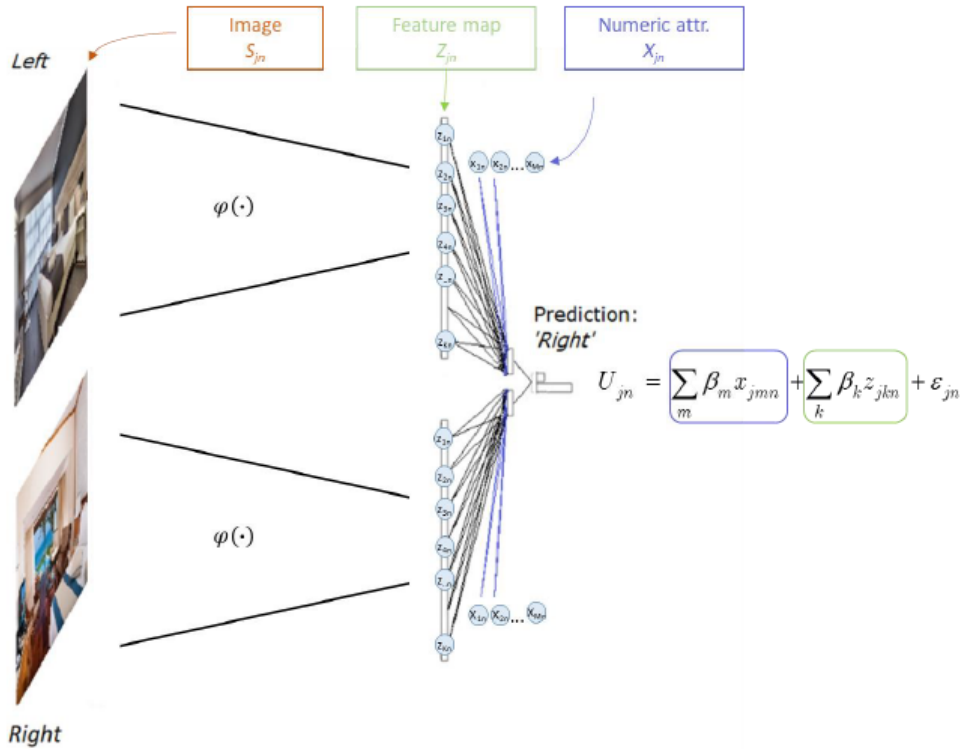
The second assumption is that utility is linear and additive with numeric attributes as well as with images' feature maps. Thus,  $f$  and  $g$  are standard linear-additive utility functions.  $Z_j = \{z_{j1}, z_{j2}, \dots, z_{jK}\}$  denotes the feature map of image  $S_j$ , and  $\varphi(w) : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^K$  is a function that maps image  $S_j$  onto feature map  $Z_j$ . Hence,  $\varphi$  is the transformation produced by the feature extractor of a CV model, and  $w$  are the associated weights (i.e., parameters). Both the numeric attributes  $X_j$  and feature map  $Z_j$  enter the utility function in a linear-additive fashion, as shown in equation 3.4. Here,  $\beta_m$  denotes the marginal utility associated with attribute  $m$ ,  $x_{jmn}$  denotes the attribute level of numeric attribute  $m$  of alternative  $j$ , as faced by decision-maker  $n$ , and  $\beta_k$  denotes the weight associated with the  $k^{\text{th}}$  element of feature map  $Z_{jn}$ . The first summation represents the utility derived from the numerical attributes, and the second summation is the utility derived from the image feature map.

$$U_{jn} = \sum_m \beta_m x_{jmn} + \sum_k \beta_k z_{jkn} + \varepsilon_{jn} \quad \text{where } Z_{jn} = \varphi(S_{jn} | w) \quad (3.4)$$

As discussed earlier, working with feature maps of an image is more effective than working with each pixel of the image. The third assumption is that  $\varepsilon_{jn}$  is independent and identically Extreme Value Type I (equation 3.5) distributed with a variance of  $\pi^2/6$ , resulting in the closed-form logit formula for the choice probabilities ( $P_{in}$ ), given in Equation 3.5, where  $C_n$  denotes the set of alternatives presented to decision maker  $n$ .

$$P_{in} = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}} \quad (3.5)$$

Figure 3.3 illustrates the model structure. A key feature of the CV-DCM architecture is the symmetrical nature of the upper and lower network components. This symmetry ensures that both the left-hand and right-hand side alternatives are treated equivalently from a mathematical standpoint, enabling the interpretation of node values in the final layer as utility measures. It is crucial to note that while we can interpret the values in the final layer as utility, this interpretability does not extend uniformly to all parameters. Specifically, while  $\beta_m$  can be understood as a marginal utility, representing the change in utility with a unit alteration in attribute level,  $\beta_k$  lacks this straightforward interpretability. The elements within the feature map,  $Z_j$ , lack clear behavioural significance or units, preventing a direct interpretation akin to that of  $\beta_m$ .



**Figure 3.3:** Model structure of the CVDCM on how the utility scores of street view images are predicted (Cranenburgh and Garrido-Valenzuela [23])

The input image is passed through a Convolutional Neural Network (CNN). This CNN extracts high-level features from the image, essentially converting the image into a numerical representation while preserving the essential characteristics of the image. These features are extracted from various layers of the CNN and capture different levels of abstraction, from simple edges and textures to complex patterns and objects. After the CNN extracts features from the image, these features are passed through a linear layer. This linear layer applies a linear transformation to the extracted features. Mathematically, this transformation can be represented as a weighted sum of the features, where the weights are learned during the training process. Once the linear transformation is applied to the image features, the model computes the total utility value. This is achieved by summing the transformed image features. Essentially, this step aggregates the importance-weighted features into a single value, which represents the estimated utility-based liveability score.

In this research, the CV-DCM model is used without the knowledge and importance of the numerical attributes. The model only takes images as input and computes the utility-based liveability score as an outcome for the specific input image.

### 3.3. XAI techniques

Law et al. [66] discusses the application of image analysis XAI for computer vision models. They stated that the need for explainability in these computer vision models for urban analytical use is increasing, while the tools for this are not yet widely known and used. They emphasize that creating an explanation with XAI is hard in complex scenarios, such as street view images, where there are multiple important objects in the image. To implement an XAI technique for explaining the model's behaviour, it is necessary to determine which explanation technique is most suited to this situation. XAI techniques can be categorized into three levels: explanation level, implementation level, and model dependency. The explanation level focuses on if the technique is used for explaining the whole model, or just a single instance/outcome of the model. If the technique is focused on the whole model, we say that the level is global, if not, then we say it is at a local level. The implementation level involves building

models with inherent interpretability (intrinsic explanations) and providing additional explanations after the model has made predictions (post hoc explanations). Intrinsic explanations involve designing machine learning models in a way that inherently provides transparency and interpretability. This could include using interpretable model architectures or incorporating features that facilitate understanding. Here, we can also think of certain model parameters and decision trees used in the model. If the model is not intrinsic, the implementation's level is post hoc. Post hoc explanations are generated after a model has made predictions/outcomes. These explanations are created specifically to clarify the model's decision-making process and make it more understandable to human users. In the model dependency level, we differentiate between techniques that are model-specific or model-agnostic explainers. Model-specific regards techniques that are only applicable to one certain type of model, and thus need readjustments when used for another type of model, while model-agnostic techniques apply to multiple types of models. In this research, the goal is to explain results from the model, therefore, a local technique is required instead of a global technique (explanation level). For the implementation level, a post hoc explainer is needed. On the model dependency level, there is no clear wish for a model agnostic or specific explainer. Since this research investigates if these XAI techniques can improve the explainability of CV-DCMs, the preference is towards the agnostic level, such that the results can be applied to further research. In conclusion, a local post hoc agnostic explainer is useful for the analysis (Adadi and Berrada [2], Murdoch et al. [97]). The reviews done in (Alicioglu and Sun [4] and Van der Velden et al. [132]) discuss multiple XAI techniques and compare them for their application to computer vision models with visual analysis. Their review is based on the medical usage of these techniques, as in this field XAI for computer vision model is used the most. From both reviews, multiple XAI techniques are recommended for use in our case. They both recommend multiple eXplainable AI techniques (LIME, SHAP (deepSHAP), LRP, and CAM (Grad-CAM)). The review on visual analysis XAI techniques done in Arrieta et al. [6] states that for post hoc explainability, LIME is the most known and applied technique in general (so not only for computer vision models, but also all other models), but also for the computer vision models discussed in this research. Due to these three reviews suggesting LIME, and it being the most used XAI technique, LIME is chosen as XAI technique for analyzing the utility-based liveability score. It is important to note here that LIME for image analysis is applied, not for other data types. LIME applies to almost every type of data input, where it works a bit differently from the image case.

### 3.4. Application of LIME for images

Local interpretable Model-agnostic explanations, LIME (Ribeiro, Singh, and Guestrin [111]), for images are used to improve the explainability and interpretability of the computer vision-enriched discrete choice model. The following paragraphs will theoretically discuss the workings of the LIME Image algorithm. All explanations are taken from the work of Garreau and Mardaoui [38].

Consider a computer-vision enriched discrete choice model  $f$  that requires an explanation of its behaviour. Additionally, assume the presence of a street-view image  $\xi$ , and the goal is to explain why this image has received a specific utility score from  $f$ . In general terms, the LIME image algorithm operates as follows:

1.  $\xi$  is decomposed into  $d$  superpixels. Superpixels are smaller homogeneous patches of the original image.
2. A new set of street view images is created  $x_1, \dots, x_n$ , which all have randomly turned off and on superpixels, resulting in a perturbed image of the original image  $\xi$ .
3. the model is then queried, where predictions  $y_i = f(x_i)$  are made.
4. a local weighted surrogate model  $\hat{\beta}_n$  is build, This simpler model fits the  $y_i$  s to this turning on and off of the superpixels.

Thus each of these superpixels of the original street view image  $\xi$  have a weighted coefficient of  $\hat{\beta}_n$ . The more positive this coefficient, the higher the importance of this superpixel is to the prediction for  $\xi$ , according to LIME. In LIME, this  $\hat{\beta}_n$  can be shown by highlighting the top positive coefficients, thereby showing the most important superpixels for the estimation. This is how LIME image works. To analyze the behaviour of LIME images further, we will discuss the superpixel creation, the sampling, the weights

and the surrogate model further. We therefore recall our computer vision enriched discrete choice model, which we called  $f$ . This function is defined on the following domain:  $f : [0, 1]^D \rightarrow \mathbb{R}$ . Again, street view image  $\xi$  is taken, where  $\xi \in [0, 1]^D$ .  $D$  is the amount of pixels of  $\xi$  where function  $f$  operates. In the street view case, the input images are 3-dimensional.

### 3.4.1. Superpixels

LIME divides  $\xi$  into superpixels. Superpixels are contiguous patches of this image that have equal colour or brightness similarities. LIME uses as default the Quickshift segmentation algorithm for creating the superpixels in the image (Vedaldi and Soatto [134]). Quickshift is a mode-seeking algorithm, which works by first estimating pixel densities in a feature space that includes colour and spatial information. Then, it identifies dense regions (called modes) and merges nearby modes to form segments. The  $k$ th superpixel of  $\xi$  is called  $J_k$ , for any  $1 \leq k \leq d$ . This results in that the  $d$  subsets  $J_1, \dots, J_d$  form a partition of the pixels.

$$J_1 \cup \dots \cup J_d = \{1, \dots, D\} \text{ and } J_k \cap J_\ell = \emptyset \forall k \neq \ell. \quad (3.6)$$

As already stated, superpixels are generally contiguous patches of the image, but this assumption is not made here.

### 3.4.2. Sampling

A fundamental concept of LIME involves generating new examples from the original image,  $\xi$ , by randomly substituting certain superpixels. This is called sampling. By default, these selected superpixels are substituted with the mean colour of the superpixel, a process referred to as mean replacement. Alternatively, it is possible to use a specific colour for replacement.

Consider a street view image  $\xi$  containing  $d$  superpixels,  $J_1, \dots, J_d$ . To sample this image, first the replacement image  $\bar{\xi} \in [0, 1]^D$  needs to be computed. If a colour  $c$  is chosen to be the colour replacing the 'off' superpixels, then  $\bar{\xi}_u = c$  for all  $1 \leq u \leq D$ . If there is no prior colour selection, then the mean image for any superpixel  $J_k$  is  $\bar{\xi} \in [0, 1]^D$ . This is defined by:

$$\forall u \in J_k, \quad \bar{\xi}_u = \frac{1}{|J_k|} \sum_{u \in J_k} \xi_u \quad (3.7)$$

For each  $1 \leq i \leq n$ , LIME randomly samples a vector  $z_i \in \{0, 1\}^d$ . This random sampling is done with an unbiased Monte Carlo strategy, i.e

$$z_i \sim B(0.5), \quad 1 \leq i \leq n \quad (3.8)$$

where  $B(p)$  is a Bernoulli-distributed random variable having probability  $p = 0.5$ . Every vector  $z_{i,j}$  corresponds to the activation ( $z_{i,j} = 1$ ) or inactivation ( $z_{i,j} = 0$ ) of superpixel  $j$ , and these  $z_i$ s are referred to as the interpretable features. More precisely, for any given  $i \in \{1, \dots, n\}$ , the new example  $x_i \in [0, 1]^D$  has pixel values given by:

$$\forall u \in J_j, \quad x_{i,u} = z_{i,j} \xi_u + (1 - z_{i,j}) \bar{\xi}_u \quad (3.9)$$

In the special case of the original image  $\xi$ , this results in the vector  $\mathbf{1} = (1, \dots, 1)^\top$ , which is just the vector representing all the superpixels are 'on'.

This random perturbing of the input images creates a synthetic neighbourhood

$$N(\xi) = \{\xi_z \mid z \in Z\} \quad (3.10)$$

### 3.4.3. Weights

The calculation for superpixel weights is done with two replacement strategies. All the new images from the neighbourhood  $N(\xi)$  are processed by the model  $f$ , resulting in the dependent variables

$$Y = \{f(\xi_x) \mid \xi_x \in N(\xi)\} \quad (3.11)$$

By random sampling, new images  $x_i$  can become a much different image than  $\xi$ . As an example, if most of the  $z_{i,j}$  are zero, then  $x_i$  is close to  $\bar{\xi}$ . The LIME image algorithm takes this into account, and new examples are given a positive weight  $\pi_i$  that takes this proximity or equality to the original image into account. For this, a distance function is adopted, in order to weight the perturbed samples differently. The intuition followed by LIME is that samples closer to  $\xi$  should weigh more. By default, these weights are defined by:

$$\forall 1 \leq i \leq n, \quad \pi_i := \exp\left(\frac{-d_{\cos}(\mathbf{1}, z_i)^2}{2\nu^2}\right) \quad (3.12)$$

Here,  $\nu > 0$  is called a positive bandwidth parameter. This parameter is equal to 0.25 by default in LIME.  $d_{\cos}$  is the cosine distance. The cosine distance can be defined as:

$$\forall u, v \in \mathbb{R}^d, \quad d_{\cos}(u, v) := 1 - \frac{u^\top v}{\|u\| \cdot \|v\|} \quad (3.13)$$

When a large amount of superpixels is 'on', then  $d_{\cos}(z_i, \mathbf{1})$  tends to become 0, and when a large amount of superpixels is turned off, this value tends to become 1. Here, the weights  $\pi_i$  are solely determined by the count of deactivated superpixels. Specifically, given that  $z_i$  has exactly  $s$  elements equal to zero, we have  $z_i^\top \mathbf{1} = d - s$  and  $\|z_i\| = \sqrt{d - s}$ . Considering  $\|\mathbf{1}\| = \sqrt{d}$  and equation 3.12, it follows that  $\pi_i = \psi(s/d)$ , where  $\psi$  is defined as:

$$\forall t \in [0, 1], \quad \psi(t) := \exp\left(\frac{-(1 - \sqrt{1 - t})^2}{2\nu^2}\right) \quad (3.14)$$

#### 3.4.4. Surrogate model

Following the initial steps, LIME proceeds to construct a surrogate model. Specifically, it constructs a linear model using the interpretable features  $z_i$  as inputs and the model predictions  $y_i := f(x_i)$  as responses. In the default setup, this linear model is derived using weighted ridge regression (Hoerl and Kennard [54]). Mathematically, the outputs of LIME for model  $f$  and image  $\xi$  are expressed as:

$$\hat{\beta}_n^\lambda \in \arg \min_{\beta \in \mathbb{R}^{d+1}} \left\{ \sum_{i=1}^n \pi_i (y_i - \beta^\top z_i)^2 + \lambda \|\beta\|^2 \right\} \quad (3.15)$$

Here,  $\lambda > 0$  serves as a regularization parameter. The coordinates of  $\hat{\beta}_n^\lambda$  are referred to as the interpretable coefficients, with the convention that the 0th coordinate of  $\hat{\beta}_n^\lambda$  represents the intercept of the model. It is worth noting that LIME typically utilizes the default setting of  $\lambda = 1$  from the sklearn library for regularization. LIME uses as a default a large  $n$ , around  $n = 1000$ , and the amount of superpixels is around 100 to 2000, then  $n \gg d$ . Then, it can be stated that  $\lambda = 0$  in the LIME image analysis. We denote the solution of equation 3.15 with  $\lambda = 0$  as  $\hat{\beta}_n$ , representing ordinary least-squares.

Now, let  $W = \{w_x \mid x \in X\}$  be the set of calculated weights per superpixel. Having the matrices of the set of masks  $X \in \{0, 1\}^{n \times k}$ , the weights  $W \in \mathbb{R}^{n \times 1}$  and the dependent variables  $Y \in \mathbb{R}^{n \times 1}$  for all the observed samples in the synthetic neighbourhood  $N(\xi)$ , then  $Y$  can be written as the response variable of the linear regression model. LIME adopts a simple linear homoscedastic model (DuMouchel and Duncan [30]) for the regression coefficients, which is

$$Y = X \cdot \beta + \epsilon \quad (3.16)$$

where the vector  $\beta$  is the weighted least squares estimator of the regression coefficients of  $Y$  on  $X$  weighted by  $W$ .

The final step of LIME for images involves displaying the superpixels associated with the top positive coefficients of  $\hat{\beta}_n^\lambda$ , which provides the user and modeller with the most influential superpixels for the model's outcome explanation, positively and negatively.

### 3.5. Segmentation metrics

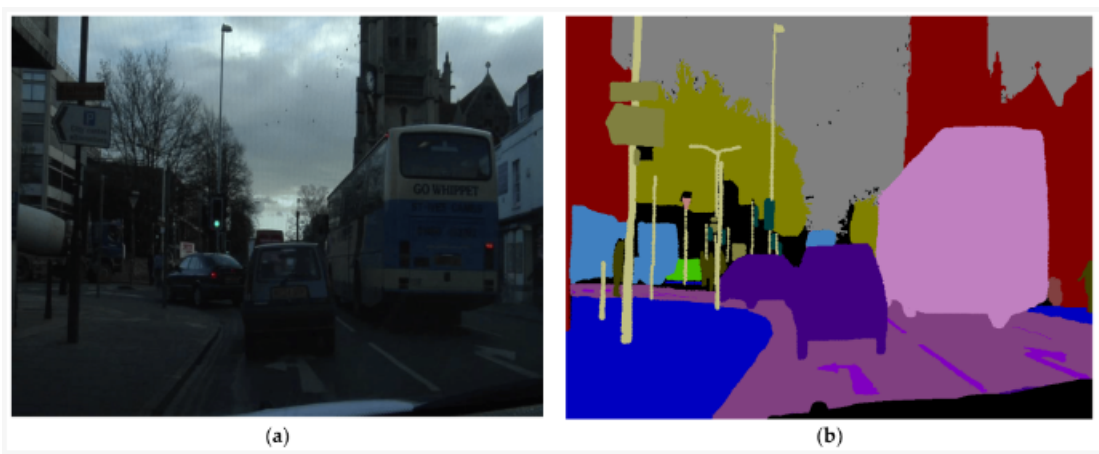
LIME works with superpixel segmentation. The segmentation algorithms chosen for this research are based on certain criteria, which are scored via a ground truth image of the original street-view image.



### 3.5.1. Ground truth images

A ground truth image is a labelled image that serves as a reference standard for evaluating the performance of segmentation algorithms. In the context of image segmentation, ground truth images contain the correct or desired segmentation labels for each pixel, indicating the true object classes or regions within the image. By comparing the segmentation algorithm's output to the ground truth, it can be assessed how accurately the algorithm identifies and segments different regions or objects in an image. Metrics like the Jaccard Index, Dice Coefficient, Adjusted Rand Index, and Variation of Information (which are discussed later on) are calculated by comparing the segmentation results with the ground truth image. These metrics provide quantitative measures of segmentation accuracy and performance. Ground truth images provide an objective basis for evaluation, allowing for consistent and repeatable assessment of different algorithms. This is crucial for scientific rigour and for comparing different methods. It allows for scientifically choosing the best fitting segmentation for the original image (Bishop [10], Gonzalez [42]).

Creating ground truth images of the original image can be done manually and automatically. Automatic segmentation algorithms like DeepLabV3 can be used for this. Figure 3.4 shows an example of a ground truth image created in DeepLabV3.



**Figure 3.4:** Example of a ground truth image, where (a) is the original street view image and (b) the ground truth image (Memon et al. [90])

In this research, ground truth images are created with LabelMe (Wada [137]), which allows the user to specify their segmentation via polygons to create a ground truth image.

### 3.5.2. Segmentation quality metrics

As discussed above, there are four metrics for analyzing segmentation quality with the ground truth image.

#### Adjusted Rand Index (ARI)

The Adjusted Rand Index (ARI) measures the agreement between two clusterings by considering all pairs of elements and counting pairs that are either assigned in the same or different clusters in the predicted and true clusterings. It adjusts for the chance grouping of elements, providing a score between -1 and 1, where 1 indicates perfect agreement, 0 indicates random labelling, and -1 indicates perfect disagreement.

This metric identifies first all pairs of elements and determines how many pairs are in the same cluster in both the predicted and ground truth segmentations and also in different clusters in both segmentations. Second, it computes the Expected Agreement, which is the expected number of such pairs for random labellings (Steinley [124]).

$$ARI = \frac{Index - Expected\ Index}{Max\ index - Expected\ Index} \quad (3.17)$$

Where the index is the number of agreeing pairs in the clusterings, the expected Index is the expected number of agreeing pairs by chance and the Max Index is the total number of pairs.

In superpixel segmentation, ARI evaluates how well the superpixels match the true segment boundaries. High ARI indicates that the algorithm correctly groups pixels that belong together in both the predicted and ground truth segmentations. The range of this value is between -1 and 1, where a good segmentation has values close to 1.

#### Jaccard Index

The Jaccard Index measures the similarity between two sets by calculating the ratio of the intersection over the union of the sets. It ranges from 0 to 1, with 1 indicating perfect overlap and 0 indicating no overlap. It is computed with the intersection and the union. The intersection is the number of pixels that are labelled the same in both segmentations and the union is the number of pixels that are labelled in at least one of the segmentations Jaccard [57].

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3.18)$$

For superpixel segmentations, the Jaccard Index measures the extent to which the predicted superpixels overlap with the true segments. A higher score means the superpixels are more accurately capturing the true segments. This metric has values between 0 and 1, where a good segmentation is indicated by values close to 1.

#### Dice

The Dice Coefficient is similar to the Jaccard Index but gives more weight to the intersection. It ranges from 0 to 1, with 1 indicating perfect agreement. It calculates the number of pixels that are labelled the same in both segmentations Dice [25].

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} \quad (3.19)$$

The Dice Coefficient evaluates the overlap between the predicted superpixels and the ground truth segments. This metric has values between 0 and 1, where a good segmentation is indicated by values close to 1.

#### Variation of Information (VoI)

Variation of Information (VoI) is an information-theoretic measure that quantifies the distance between two clusterings. It measures the amount of information lost and gained in changing from one clustering to the other. It computes the entropy of each clustering, representing the uncertainty, after which it calculates the mutual information between the two clusterings. VoI assesses how much information is shared between the predicted and ground truth segmentations. Lower VoI values indicate better segmentation performance, as less information is lost or gained (Meila [89]).

$$\text{VoI} = H(A) + H(B) - 2I(A; B) \quad (3.20)$$

Here,  $H(A)$  and  $H(B)$  are the entropies of clusterings A and B, and  $I(A; B)$  is the mutual information.  $H(B; A)$  is the under-segmentation error, this component measures the amount of information lost when the ground truth segmentation B is used to describe the predicted segmentation A. It captures how much of the ground truth's detail is missed by the predicted segmentation.  $H(A; B)$  is the over-segmentation error. This component measures the amount of extra information introduced by the predicted segmentation A when describing the ground truth segmentation B. It captures how much the predicted segmentation splits the regions of the ground truth. Over and under-segmentation can generally be written as follows:

$$H(X; Y) = H(X) - I(X; Y) \quad (3.21)$$

With this, we can see that the  $\text{VoI}$  formula can also be written as  $H(B; A) + H(A; B)$ , the sum of the over and under-segmentation errors.

Interpretation of these scores is needed to understand their implications. A high ARI suggests that the superpixel segmentation closely matches the true segments, accurately grouping related pixels. A high Jaccard Index says that there is a significant overlap between the predicted and true segments, indicating good segmentation. A high Dice Coefficient suggests that the segmentation algorithm has a high level of agreement with the ground truth, correctly identifying segment boundaries. A low Vol states that there is minimal difference in information between the predicted and true segmentations, suggesting accurate segmentation. This metric quantifies the difference by measuring both under-segmentation and over-segmentation errors. These two components provide a more detailed analysis of the discrepancies between the segmentations, which is why Vol can return two values. This metric has values ranging from 0 to infinity, where values close to zero mean a good segmentation. In practical implementations, it is often useful to look at both the under and over-segmentation error separately to understand how the predicted segmentation deviates from the ground truth. A high under-segmentation error might indicate that the predicted segmentation fails to capture the fine details of the ground truth and a high over-segmentation error might indicate that the predicted segmentation is too fragmented compared to the ground truth.

### 3.6. LIME metrics

Three measures are used to evaluate whether the created LIME explanation is of sufficient quality: the binary class ratio, the probability distribution uniformity metric, and the coefficient of covariance.

#### 3.6.1. Binary class ratio

As already discussed, a good distribution between the switching of the predicted classes (in our case binary) is needed for LIME to build a good linear model. Therefore, the ratio of the class 'good' and 'bad' is important. This ratio should be around 0.5, as this would mean that around half of the samples are classified as 'good' and half as 'bad'. This ratio is the percentage of equal classifications as the original image of the total amount of perturbed images/samples called the Binary Class Ratio (BCR) (Rashid et al. [110]). A good BCR value is around 0.5.

$$BCR = \frac{\text{Total amount of equally classified samples}}{\text{Total amount of samples}} \quad (3.22)$$

#### 3.6.2. Probability Distribution Uniformity Metric

The Probability Distribution Uniformity Metric (PDUM) evaluates the distribution of probabilistic classifications in a binary classification model. In the context of classifying street views into "good" and "bad" categories, the PDUM assesses whether the probabilities assigned to each classification are evenly distributed across the range from 0 to 1.

A well-distributed set of probabilities ensures that the model does not produce overly confident predictions, allowing for a more nuanced and interpretable output. For effective use of interpretative tools like LIME, it is beneficial to have probabilities that are not skewed towards 0 or 1, as a uniform distribution enables more robust analysis and reliable explanations. A uniform probability distribution means that the model expresses varying degrees of confidence in the predictions of the model (Rashid et al. [110]).

By visual analysis of the probability classification distribution, this metric will be evaluated per street view.

#### 3.6.3. Coefficient of covariance

When situations occur where the distribution of these classifications is problematic, the  $\beta$  vector's values drop to small values. This results in the variability decreased between the feature importance values. Rashid et al. [110] calls this 'confusion'. If the weights become too small, or the variation between them is too small, their information loses value. To measure this, the standard coefficient of variation is used, CV.

$$CV(\beta) = \frac{\sigma_{\beta}}{\mu_{\beta}}, \quad (3.23)$$

where  $\sigma_{\beta}$  and  $\mu_{\beta}$  are the standard deviations and the mean of the beta values. In this context,  $\sigma_{\beta}$  and  $\mu_{\beta}$  denote the standard deviation and mean of  $\beta$ , respectively. For a good  $CV(\beta)$ , the value should not

approach zero. A near-zero  $CV(\beta)$  would suggest that all superpixels possess nearly the same value, making it difficult to distinguish distinct sub-regions within the image.

These metrics answer the fifth sub-question: 'How can the performance of LIME be measured?'.

The methodological framework established in this chapter sets the stage for the practical application described in the next chapter, which will detail the specific implementation steps taken to apply our chosen methods.

# 4

## Implementation

This chapter details the technical execution of the proposed methodology. The research design for this phase is practical and experimental, focusing on the application and validation of the chosen methodologies.

### 4.1. Segmentation algorithm choice

As discussed in the methodology, the first important choice for the application of the LIME algorithm is how the superpixels are chosen. This is done by the segmentation algorithm and directly influences how the explanation of LIME is created. This segmentation algorithm is desired to create a semantic meaningful segmentation. Semantic meaning refers to the meaning derived from the content of an image in a way that is understandable and relevant to humans. In the context of image processing and analysis, it involves interpreting the visual content to recognize and label objects, scenes, and other features in a way that aligns with human understanding and context. For example, in street view images, semantic meaning could involve identifying and understanding elements such as cars, pedestrians, road signs, buildings, and other environmental features. This allows the AI system to provide explanations that are not just statistically significant but also meaningful and interpretable by humans. Incorporating semantic meaning in LIME Image explanations means ensuring that the explanations provided by LIME are aligned with human interpretable features. It also thus allows for that no prior object detection is used, but that the explanations are understandable due to the possibility of object detection in these explanations. A clear segmentation sort of 'performs' as this object detector. This involves segmenting the image into meaningful parts and using these segments to explain the model's decisions. Here, multiple different superpixel segmentation algorithms are possible. To increase the explainability of the model's decision behaviour, a segmentation algorithm is desired which provides the street view image with superpixels that are semantically meaningful. First, superpixels with semantic meaning ensure that the segmented parts of the image correspond to recognizable and meaningful features, such as trees, sidewalks, buildings, or parks. This alignment is crucial because it allows for a more accurate interpretation of the model's decisions. Second, having semantically meaningful superpixels enhances human interpretability. Explanations become clearer and more understandable when the perturbed image regions correspond to recognisable objects or areas. Third, good segmentation leads to visual coherence in the explanations. Superpixels that follow semantic boundaries result in more realistic perturbations. Thus, selecting a good segmentation algorithm that produces superpixels with semantic meaning is crucial for using LIME as it enhances the interpretability of the explanations.

Relating this to the perceived liveability literature review, incorporating semantic meaning in superpixel segmentation allows the analysis to connect model results with important features identified in the literature on perceived liveability. The perceived liveability literature review highlights various elements in street views, such as greenery, building conditions, and physical incivilities, which significantly impact residents' satisfaction with their living environment. By ensuring that the superpixels generated by the segmentation algorithm are semantically meaningful, we can more accurately link these segments to

the features that are crucial for perceived liveability.

By aligning the segmentation process with the perceived liveability factors, the model's decisions can be grounded in real-world features that are known to influence human perceptions of liveability. This approach enhances the overall reliability and interpretability of the model, making it a more effective tool for urban planning and policy-making aimed at improving liveability in urban environments.

Relating to the XAI literature review, two primary aspects are crucial: policy-related evaluation focused on interpretability and explainability, and the scientific evaluation of the quality of explanations. The policy-related aspect emphasizes the need for AI systems to be transparent and understandable to non-expert stakeholders, including policymakers and the general public. This transparency is essential for building trust, ensuring accountability, and facilitating informed decision-making based on AI-generated insights. Therefore, explanations should be clear and comprehensible to users. This means that the superpixels or features highlighted by LIME must correspond to elements that users can recognize and understand. The explanations should accurately reflect the model's decision-making processes. This means that the features identified by LIME as important should genuinely influence the model's predictions. In the context of urban liveability, coherence in explanations is vital. This involves ensuring that the segments identified by LIME are consistent with the meaningful features discussed in the perceived liveability literature. For example, explanations should consistently highlight features such as greenery, building conditions, and physical incivilities, which are known to impact perceived liveability. This coherence enhances the reliability and relevance of the explanations, making them more useful for urban planners and policymakers. Here, the correct choice for the segmentation algorithm affects these outcomes heavily.

LIME can use all segmentation algorithms from the Python package *skimage.segmentation*. For this research, seven segmentation algorithms have been analyzed for segmenting street view images: Quickshift, Felzenszwalb, SLIC, Compact Watershed (as also discussed in (Schallner et al. [118]) Mean Shift, SLICO, and ERS. For further analysis, only Quickshift, Felzenszwalb and SLIC are deemed to be best fitted for segmenting street view images in semantic meaningful segmentation (as also discussed in (Schallner et al. [118])). In Appendix H the other four segmentation algorithms are shortly discussed and shown.

Quickshift is an algorithm for image segmentation that is based on kernel density estimation. It is a non-parametric method that doesn't require prior knowledge of the number of segments in the image. Quickshift works by iteratively assigning pixels to nearby clusters based on the similarity of colour and texture features. It essentially identifies dense regions in feature space, which correspond to homogeneous regions in the image. It has the advantage that it doesn't require prior specification of the number of segments, has fast computation time, especially for small to medium-sized images, and can handle images with varying scales and resolutions (Vedaldi and Soatto [134]). It uses the following parameters which can be changed:

- Kernel size  
The kernel size defines the spatial window over which density estimation is performed. The kernel size determines how much influence a pixel's neighbourhood has on the image clustering. A small Kernel leads to more detailed segmentation with smaller and more numerous segments, which is useful for capturing fine details but can result in noisy segments. A large Kernel size produces coarser segmentation with larger, smoother segments.
- Maximum Distance  
The maximum distance sets the threshold for merging clusters. It controls the maximum distance in the feature space (including colour and spatial coordinates) within which pixels can be considered part of the same segment. A small Maximum Distance leads to more segments, as pixels need to be very close in the feature space to be merged. A large Maximum Distance results in fewer segments, as pixels can be farther apart in the feature space and still be merged.
- Ratio  
The ratio parameter controls the balance between colour and spatial proximity. It determines the importance of colour similarity relative to spatial proximity when merging pixels into segments.

A low Ratio emphasizes spatial proximity over colour similarity. Pixels that are close together spatially will be more likely to be merged, even if their colours are different. This can result in segments that follow the spatial structure but might mix different colours. A high Ratio emphasizes colour similarity over spatial proximity. Pixels with similar colours will be more likely to be merged, even if they are farther apart spatially. This can result in segments that are more color-consistent but might ignore spatial coherence.

Felzenszwalb agglomerates pixels on a graph by defining a metric to measure the evidence for a boundary between two regions using a graph-based representation of the image. Superpixels are generated by finding the minimum spanning tree of the constituent pixels. Dijkstra's algorithm is used to compute the shortest paths in the undirected graph defined on these grid positions. The resulting superpixels have irregular sizes and shapes and it cannot directly control the number of superpixels (Felzenszwalb and Huttenlocher [36]).

- **Scale**  
Scale controls the size of the segments by influencing the threshold function that determines if regions should be merged based on the difference in intensity. A higher scale means more significant intensity differences are required for a boundary to be considered, leading to larger segments.
- **Sigma**  
The sigma parameter is the standard deviation of the Gaussian smoothing applied to the image before segmentation. This helps to remove noise and minor details that could otherwise lead to over-segmentation. A small Sigma means minimal smoothing, preserving more details and potential noise. This is useful for high-detail segmentation but can result in noisy segments. A large Sigma gives more smoothing, reducing noise and minor details, which is useful for more robust segmentation but might smooth out important details.
- **Minimum Size**  
The minimum size parameter sets the smallest allowable size for segments. Segments smaller than this size will be merged with neighbouring segments, ensuring that all final segments are at least this large.

The SLIC (Simple Linear Iterative Clustering) algorithm begins by randomly initializing cluster centres (K-Means clustering), followed by an assignment step where each pixel is associated with the nearest cluster centre whose search region overlaps the pixels location. After this initial assignment, an update step adjusts the cluster centres based on the associated pixels. These assignment and update steps are repeated iteratively until the error converges. Finally, a post-processing step enforces connectivity by reassigning disjoint pixels to nearby superpixels, resulting in regular superpixels. SLIC superpixels adhere well and efficiently to image boundaries. However, the superpixels generated by SLIC cannot capture global image properties (Achanta et al. [1]).

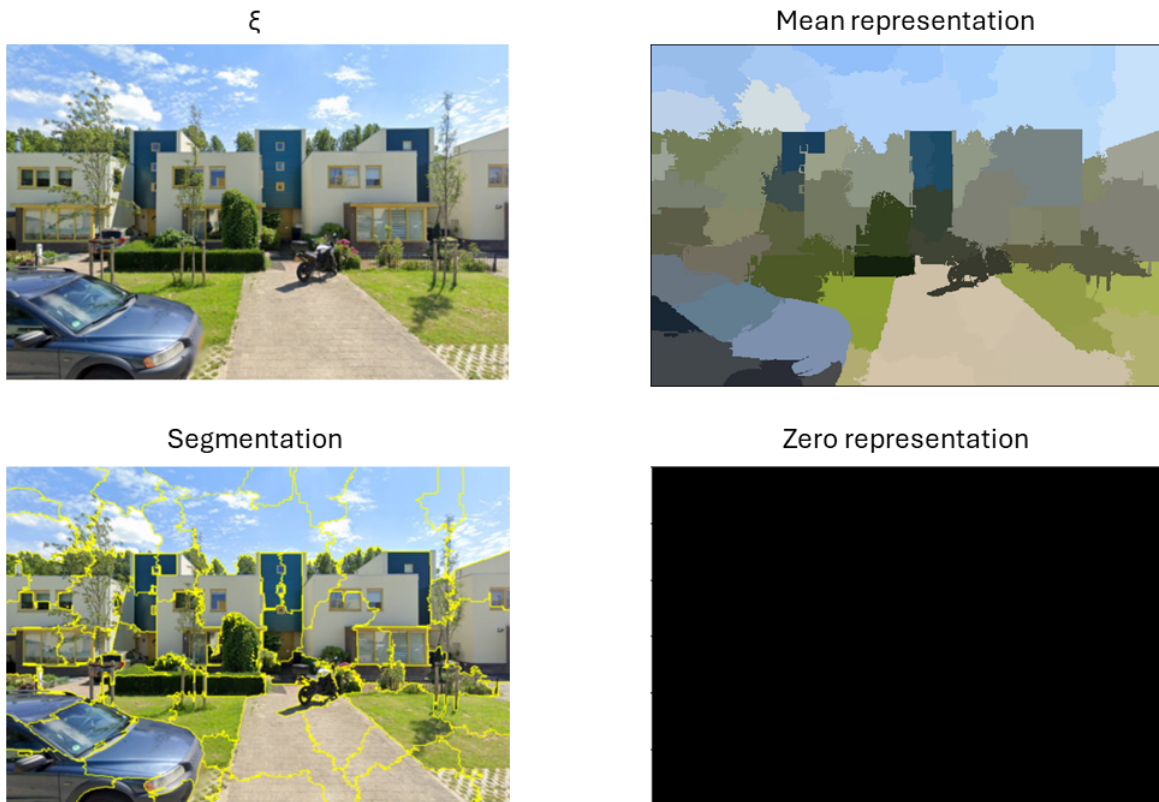
- **Number of segments**  
The number of segments parameter defines the approximate number of superpixels (segments) that the algorithm should produce. The smaller the number of segments, the larger the superpixels.
- **Compactness**  
The compactness parameter balances the importance of colour similarity versus spatial proximity. A higher compactness value makes the algorithm prioritize spatial proximity more, leading to more compact and evenly shaped superpixels. A low Compactness prioritizes colour similarity, allowing superpixels to be more irregular in shape and spread out spatially, whereas high compactness emphasizes spatial proximity, producing more regularly shaped and spatially compact superpixels.

Figure 4.1 shows these three segmentation algorithms with random parameter values being used on one specific street view image.



**Figure 4.1:** Segmentation algorithm comparison: Felzenszwalb's (scale=1200, sigma = 0.5, min size = 50), SLIC (segments = 50, compactnes=10), Quickshift (Kernel size = 12)

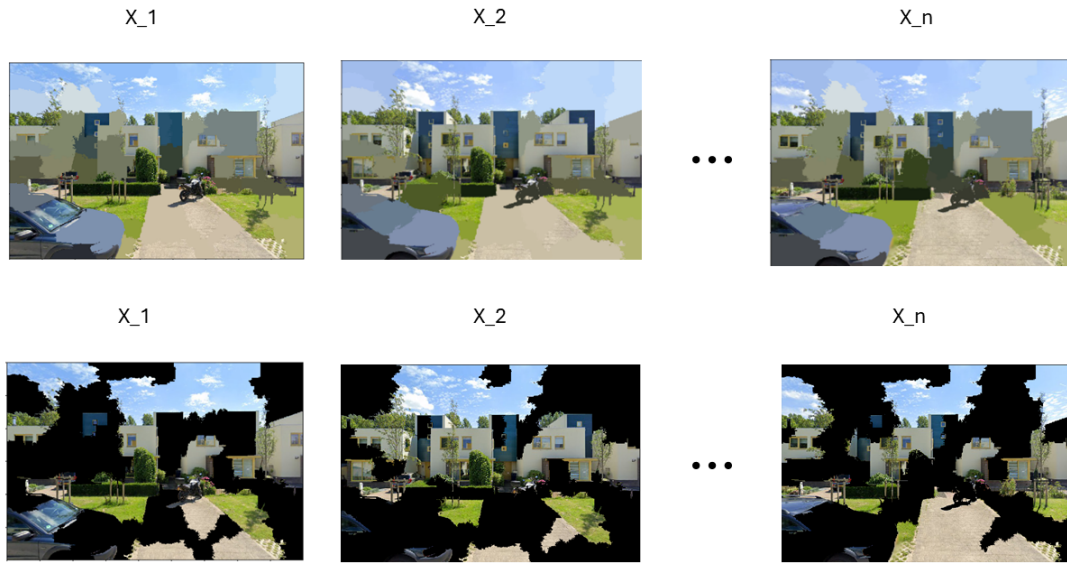
To further explain how the choice of segmentation alters the LIME explanations, the image superpixel creation and perturbation behaviour are shown for this street view image. This image,  $\xi$ , which has the perturbed images  $x_1, \dots, x_n$ . First, the segmentation is made and the mean representation is shown, along with the zero representation where all superpixels are turned off and turned black.



**Figure 4.2:** From left to right: the original street view image, its mean representation (each superpixel replaced by its mean colour), the segmented street view image, the zero representation (each superpixel replaced by the chosen colour black).

Then, it creates perturbations which are used to assess the weights of the linear model for each superpixel.





**Figure 4.3:** Perturbed images of the original image, top row with mean replacement and the bottom row with colour black as a replacement.

This shows the importance of the segmentation algorithm, as the LIME algorithm itself is highly dependent on the segmentation quality. Choosing a good segmentation algorithm for street view images is critical because of the inherent complexity and variety present in such images.

## 4.2. Classification problem

For LIME to generate explanations, it needs to work with a complex model that provides a clear decision boundary. Because LIME creates explanations by generating a local, interpretable model around a specific prediction, it needs to understand how the model's decision boundary behaves near the prediction. Classification models inherently have decision boundaries that separate different classes, making it easier for LIME to create a local linear approximation. Classification models produce binary (or multi-class) outcomes that simplify the interpretability process. Regression models, which predict continuous outcomes, do not have clear decision boundaries, making it not possible for LIME to generate simple, interpretable explanations for image analysis. In the case of this research, the classification is binary and divided into images being labelled as either 'good' or 'bad'. The local decision boundary therefore is well-defined, allowing LIME to fit a linear model that approximates the classification decision boundary.

The complex model used in this research (Cranenburgh and Garrido-Valenzuela [23]) is a regression model. By converting a regression model to a classification model using a threshold, a binary outcome model is created that LIME can use to understand how different features influence the prediction. Equation 4.1 shows the simplified classification model.

$$\text{Classification} = \begin{cases} \text{'Good'}, & \text{if utility score} \geq \text{threshold} \\ \text{'Bad'} & \text{otherwise} \end{cases} \quad (4.1)$$

Here, the threshold is decided by the modeller. This threshold is specific to the input image. In the next paragraph, more information about this is given. This threshold simplifies the interpretation, allowing LIME to explain whether changes in features push the prediction towards "good" or "bad". The threshold defines what is considered "good" or "bad". Changing this threshold alters the decision boundary, which in turn affects LIME's local linear approximation. A different threshold can change the features that LIME identifies as important, as it changes the classification outcome for perturbed samples. If

the threshold is too high or too low, it might not align well with the underlying distribution of perturbed image utility scores, leading to inconsistent or less meaningful explanations. The linear model in LIME is trained on the perturbed samples and their corresponding predictions. This training process helps LIME identify which features are most influential in the local decision-making process. If the perturbed samples change the classification outcome (e.g., from "good" to "bad"), LIME can learn which features contribute to these changes, providing insight into the original model's behaviour. Therefore, the quality of the perturbation sampling and the choice of threshold affect the accuracy of the local linear model. Effective perturbation sampling ensures that the local approximation accurately reflects the complex model's behaviour in the neighbourhood of the original prediction. If perturbations are not representative or if the threshold is poorly chosen, the linear model might not accurately capture the decision boundary, leading to misleading explanations.

In this research, two different classification algorithms are employed: Deterministic Binary Classification and Probabilistic Binary Classification. Analyzing both classification methods is useful as it allows the exploration of two distinct analytical approaches to the problem, the deterministic way, which provides clear, binary outcomes, and the probabilistic way, which offers a spectrum of probabilities for each classification decision. This dual approach can reveal different insights and nuances about the data and the model's behaviour. Also, by comparing the outcomes of deterministic and probabilistic classifications, the research can evaluate whether the choice of classification method significantly influences the results. Lastly, employing and comparing both methods adds robustness to the research findings. It ensures that the conclusions drawn are not dependent on a single classification approach, thereby increasing the reliability and generalizability of the results.

With deterministic binary classification, the classification into 'good' or 'bad' is based on strict thresholds with probabilities of 100% or 0%. If the utility score of an image meets or exceeds the threshold, it is classified as 'good'; otherwise, it is classified as 'bad'. This method provides a clear decision boundary, and for policy advice creation, the deterministic approach's simplicity and clarity can be advantageous, as it provides clear-cut decisions that are easy to communicate to stakeholders and in theory should be easier to use for explanation on what makes a street view good or bad. However, the classifications rigidity might overlook nuanced information, potentially leading to less flexible policy recommendations that fail to capture gradual shifts in data patterns, as small changes near the threshold can result in different classifications, which might not reflect gradual changes in utility.

The probabilistic binary classification is based on probabilistic thresholds that introduce gradations in the classification probabilities. The model uses thresholds, but instead of strict binary outputs, it provides probabilities that an image is 'good' or 'bad' based on their respective utility score. The classification probabilities are influenced by three key thresholds, instead of the one threshold in the deterministic classification: a lower threshold, an upper threshold, and a mid-threshold. If an image's utility score is below the lower threshold, it is classified as 'bad' with 100% probability and 0% 'good'. If the score is above the upper threshold, it is classified as 'good' with 100% probability and 0% 'bad'. For utility scores between these thresholds, the probabilities are calculated proportionally, reflecting a more nuanced transition from 'bad' to 'good'. This means that at the middle threshold, the probability for the 'bad' and 'good' classification is 50% for both. In between this lower threshold and the upper threshold, the probability for the classification 'bad' alters proportionally from 100% to 0% respectively, and for the classification the opposite holds.

This probabilistic approach offers flexibility and can better represent gradual changes in utility scores, providing a more detailed understanding of the model's behaviour. However, it is more complex to interpret compared to the deterministic binary model, and the probabilistic nature can introduce uncertainty in classification decisions.

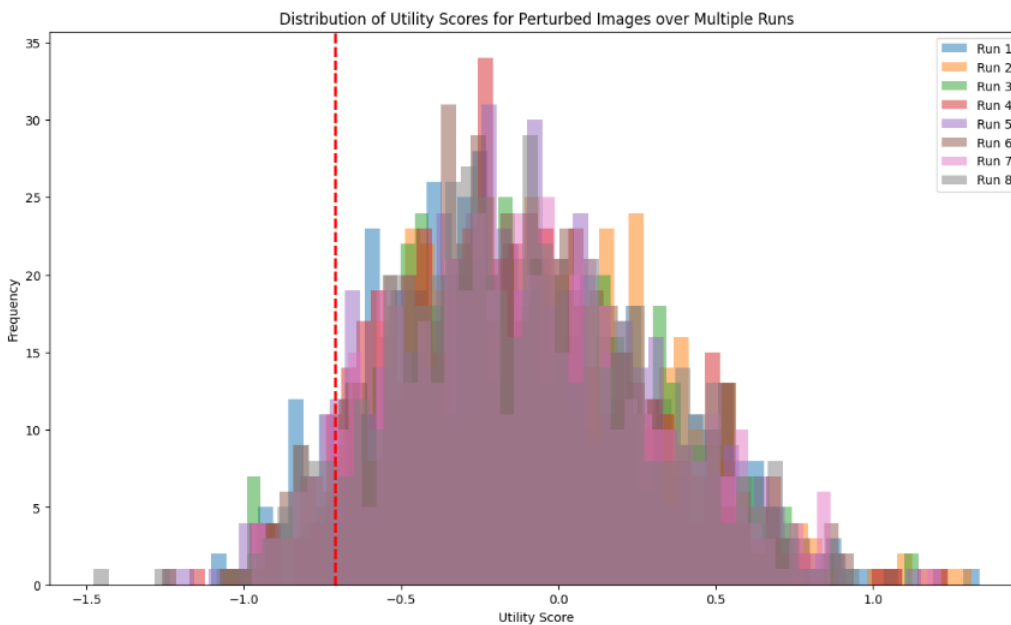
For policy advice creation, the probabilistic approach's flexibility is in theory beneficial as it allows for more nuanced recommendations that account for varying degrees of certainty. This can help policy-makers make more informed decisions by considering a broader range of scenarios.

In summary, while deterministic binary classification provides a simpler threshold, and therefore the

simpler classification of when an image is good or bad, probabilistic classification offers a more nuanced and flexible understanding of the model's behaviour and allows us to account for the subtle utility differences in classification. Both approaches have their strengths and weaknesses, and the choice between them should consider the specific needs of the research and the nature of the data being analyzed. To analyze if one of the two classification methods is beneficial and thus optimizes the enhanced explainability of the model, both will be analyzed in case studies.

### 4.3. Perturbation utility score changes

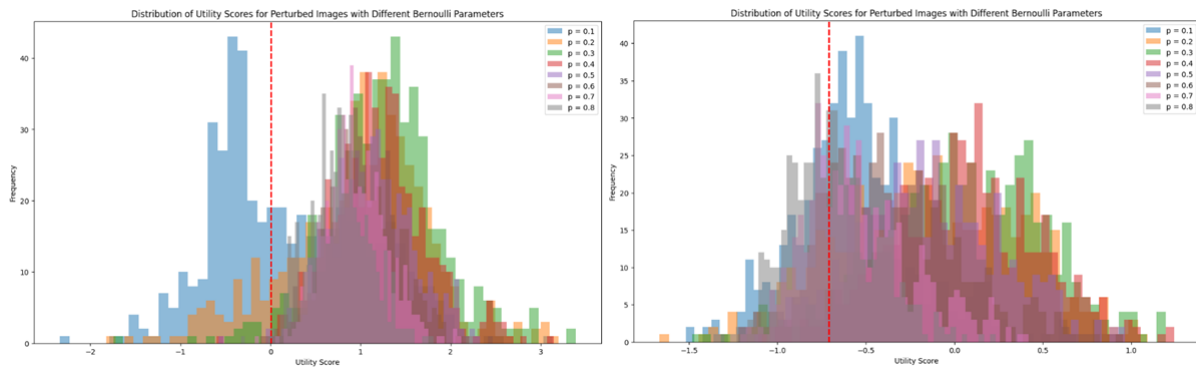
To determine a valid classification threshold value, a good understanding of the distribution of the perturbed image utility scores is needed. The sampling process of the perturbed images is done with a Bernoulli distribution with a parameter value of 0.5. This results in practice that a large part (on average 50 per cent) of the superpixels are turned off per perturbation in the sampling process. This means that the perturbed images are quite distinct from the original image (as could already be seen in Figure 4.3), and therefore, the utility score of the perturbed images can be quite different from that of the original image. Figure 4.4 shows a sampling of a street view image with the Quickshift algorithm. Here, 8 runs are done for robustness and each run, the utility score distribution of the perturbed images is shown. The red dotted line indicates the utility score of the original image which is sampled.



**Figure 4.4:** Distribution of the utility scores of the perturbed images of the original image on 8 different model runs.

As can be seen, the perturbed images score on average much different from the original image. This behaviour is seen in all street view images. This is an important observation, as this utility score is used for the classification threshold in both classification algorithms. In the deterministic classification, this is important, while in the probabilistic due to the evenly distributed probabilities, this is less of a problem. Thus especially in the deterministic classification, the behaviour of the perturbed images needs to be analyzed. If a threshold is chosen which is near the red dotted line, then there are fewer classifications which switch from 'good' to 'bad' or vice versa, which is needed for LIME to fit the linear model correctly. Since the deterministic classification in this research is binary ('good' or 'bad'), the perfect distribution of these perturbed image classifications is 50/50 (per X amount of samples created). Taking the mean of this utility score distribution of the perturbed images is a valid threshold.

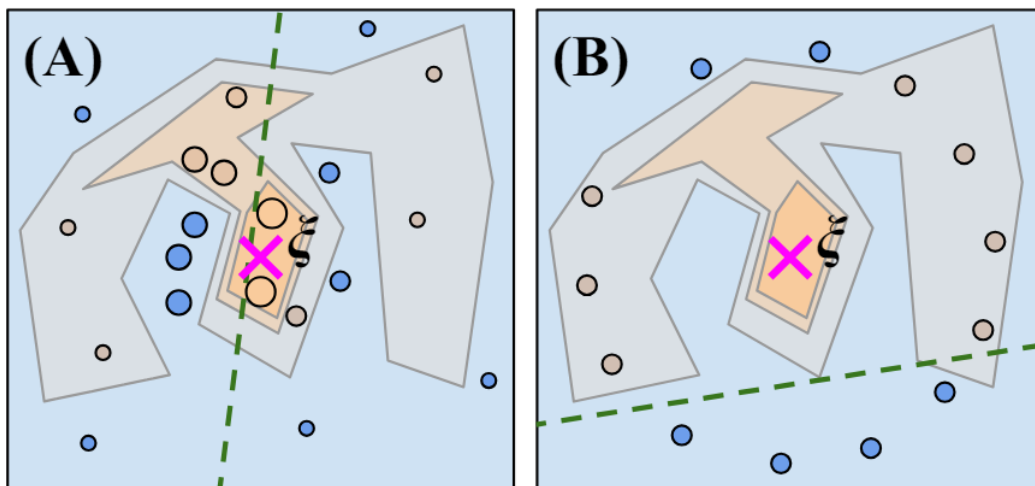
The results above are all based on the parameter value of the Bernoulli distribution equalling 0.5. In figure 4.5 the effects of the Bernoulli parameter are analyzed, where different parameter values are taken to observe the effects.



**Figure 4.5:** Bernoulli parameter influence analysis. Two different street view images with the distribution of the utility scores of the 500 samples/perturbed images per Bernoulli parameter choice.

As can be seen, the choice for the Bernoulli parameter is important. For different values of the Bernoulli parameter, the distribution of the utility scores of the perturbed image changes as well, but as can be seen for most parameter values, the same behaviour occurs in computing more extreme utility scores than the original image. Since for each parameter value this behaviour is seen, the decision is made to remain with using the 0.5 Bernoulli parameter value for sampling.

This behaviour of LIME is possibly due to one significant limitation of the sampling process, the under-representation of the neighbourhood of the explained sample,  $\xi$ . In creating the explanation, LIME creates a synthetic neighbourhood around  $\xi$ ,  $N(\xi)$ . This approach works best when the number of superpixels is small. The masks for the perturbed samples are generated via the Bernoulli coefficient of 0.5, meaning the probability of a mask containing a specific number of active superpixels follows a binomial distribution, where the probability mass function gives that the likelihood of sampling masks with high or low numbers of active superpixels is very low. Therefore, the samples tend to cluster around configurations with approximately 50 per cent of superpixels masked. This results in LIME Image creating a hypersphere of samples forms around the original images, instead of samples that are very close to it. This results in an under-representation of the true local behaviour of the model around  $\xi$ . This can be seen in figure 4.6. Therefore, while LIME aims to be locally faithful, the inherent limitations of the sampling process can undermine LIME's effectiveness, especially in complex image scenarios (Vermeire et al. [135], Rashid et al. [110]).



**Figure 4.6:** How LIME is supposed to work (A) with samples close to the original image, and how it works (B) using Monte Carlo sampling (Rashid et al. [110])

This problem of altering perturbation scores together with the complex nature of street view images

makes it hard to provide correct segmentation answers to the fourth sub-question: 'What challenges arise with the application of Street View Images for LIME image analysis?'

## 4.4. LIME Explanation visualization

The programming language Python has a package for the implementation of LIME, for every data type. For images, the explanation object is called Image Explainer in the Python Package. This package builds the explanation given the parameters and functions earlier discussed in this chapter. An important note is that it can be chosen by the modeller for how many classes this explainer can be built. In this, research, only two classes are present. Therefore, LIME can create explanations for the classification 'bad', but also 'good' for each street view image. It has this ability because in the classification algorithms, each perturbed image has a probability for 'bad' and 'good', and therefore the explanation can be built for both. The explanation can be visualized to understand the explanation. These explanations are all based on the weights of the superpixels, which first need to be clarified further.

The weights represent the importance or influence of each superpixel on the model's prediction. The weight assigned to a superpixel indicates how much that particular segment of the image contributed to the model's prediction. Positive weights suggest that the superpixel contributed positively towards the predicted class, while negative weights indicate a negative contribution. Superpixels with positive weights are those that push the model towards predicting the current class. These regions contain features that the model associates with the predicted label. In our research, the positive weights indicate that the superpixel contributes to the image scoring higher than the threshold, and the negative contributes to scoring lower than the threshold. The absolute value of the weight indicates the strength of the influence. Larger absolute values mean stronger influence, regardless of whether it is positive or negative. This understanding is used in the three visualization techniques.

### 4.4.1. Most important superpixels

One visualization option is to show the  $X$  amount of the most important superpixels. In this representation, the image is shown with the segmentation superpixels. The  $X$  most influential superpixels (so having the absolute most weight), are highlighted. If these superpixels have a positive weight, they are showcased in green, and otherwise in red. This helps users identify which parts of the image had the greatest impact on the model's prediction, offering a clear and concise way to understand the model's decision behaviour.

### 4.4.2. Hide function

The visualisation with the LIME hide function allows the user to hide all the other superpixels which are not in the chosen set of the user. By isolating either the positively contributing superpixels or the negatively contributing ones, users can better understand how different parts of the image support or contradict the model's prediction. This binary separation helps in further analyzing the specific features that are driving the model's decisions in either direction.

### 4.4.3. Heatmap

A heatmap highlights the importance or influence of each superpixel on the model's decision. In this visualization, each superpixel is coloured according to the superpixel weight, with different colours or intensities representing the strength and direction of the contribution. The heatmap provides a more nuanced view compared to binary highlighting, as it allows users to see the gradient of influence across the entire image. This can be particularly insightful for understanding subtle nuances in how the model interprets various parts of the image.

## 4.5. Interpretation of LIME

The application of the LIME algorithm within this model is designed to enhance the interpretability and transparency of the machine learning processes used for assessing the utility-based liveability score. LIME generates explanations for individual predictions by approximating the model locally with a simpler, interpretable model. This means that for each street view image, LIME can identify which parts of the image (superpixels) are most influential in classifying the image as contributing to good or bad

liveability. The use of a segmentation algorithm that provides semantically meaningful superpixels ensures that the explanations generated by LIME correspond to recognizable features in the images, such as the ones discussed in the literature review. This alignment with human-interpretable features enhances the usefulness of the explanations. By incorporating insights from this literature review on perceived liveability, the LIME model's explanations can be related to known important features such as greenery, building conditions, and physical incivilities for example. This allows for a more comprehensive understanding of the model's outputs in the context of established urban liveability research. The LIME model can show how different visual elements in street view images, such as these features, influence the classification of an area as having good or bad liveability and show the relation they have to each other. The explanations provided by LIME can be checked against the factors identified in the literature on perceived liveability. LIME thus helps in understanding the decision-making process of the CVDCM model by revealing which parts of the image contribute most to the model's predictions. This transparency is crucial for evaluating the model's reliability and trustworthiness.

The explanations generated by the LIME model are interpreted differently depending on the chosen classification algorithm—deterministic or probabilistic. This differentiation stems from the use of a threshold in our regression-based model, which impacts how LIME explanations are framed compared to standard LIME applications.

In the deterministic classification, a unique threshold is chosen for each image to determine the binary classification for the image. For instance, if an original image scores -1 and the threshold is -0.2, the image is classified as 'bad'. LIME explains why the image falls in the 'bad' category by considering the interval  $(-\infty, -0.2)$ . In this context, 'bad' represents this entire interval rather than just the score of -1. Positive weights in the LIME explanation indicate superpixels contributing to the 'bad' classification, meaning these regions push the score further below -0.2. Conversely, negative weights indicate superpixels that would contribute to a 'good' classification by pushing the score towards or above the threshold, thus potentially reclassifying the image as 'good'. If an image is classified as 'good', the opposite interpretation applies. Each LIME explanation in the deterministic classification is thus unique to the original image and the image's specific classification based on the image's utility score and the chosen threshold. The weights' influence on classification can vary with different thresholds, highlighting different features for each classification scenario.

In the probabilistic classification, the classification of 'bad' or 'good' is determined by probabilistic values. Here, negatively weighted superpixels decrease the probability of the image being classified as the target category, while positively weighted superpixels increase this probability. Unlike the deterministic approach, the probabilistic classification does not require a unique threshold for each image. Instead, a single threshold can be applied across all images due to the nature of probability-based classification. This probabilistic interpretation allows for a more nuanced understanding of how each superpixel influences the likelihood of the image falling into the 'good' or 'bad' category, providing a gradient of influence rather than a binary distinction.

## 4.6. Implementation pipeline

A pipeline for the usage of LIME for street-view images can be created to evaluate the explanation and ensure that the build explanation is valid.

1. Analyzing the segmentation algorithms visually to find the best segmentation algorithm
2. Create ground truth image and apply the metrics for the segmentation algorithms
3. Decision based on both steps to choose optimal segmentation algorithm
4. Classification analysis on which threshold to choose
5. Create LIME explanation for both classifications
6. Showcasing results with the three visualization techniques
7. Analyzing the LIME metrics

# 5

## System and Actor analysis of human perception modelling at the municipality of Rotterdam

In the preceding chapters, a robust methodological framework is established. Building on this foundation, the current chapter delves into a system and actor analysis to explore how and where the modelling of preferred liveability, augmented with LIME, aligns with the concerns and operations of the municipality of Rotterdam in managing public spaces. By examining the intricate network of stakeholders, the aim is to identify the key actors influencing and being influenced by urban liveability metrics. This analysis will illuminate how the methodology integrates with the municipality's existing operations, ensuring that this research not only aligns with but also enhances the ongoing efforts to optimize public spaces for improved liveability. Therefore, a system and actor analysis are discussed to analyze where and how human perception AI models can be incorporated into the decision-making process of the municipality of Rotterdam. The scope of decision-making in the public space is on the municipal level. A system analysis is a method used to understand the broader context and components of a system, including the environment, interactions, and feedback loops (Meadows [88]). An actor analysis, on the other hand, identifies and examines the key stakeholders, their roles, interests, and influence within a system (Bryson [12]).

First, we will discuss the system that the municipality of Rotterdam is using to model the human perception of public space. Second, the actor analysis will focus on the municipality of Rotterdam, detailing how various actors are concerned with and influence the human perception of public space.

### 5.1. System analysis

Municipal governments in the Netherlands are crucial actors in the country's public administration, responsible for a wide range of services and regulations that directly affect local citizens (Rijksoverheid [113]). The structure of municipal governments typically includes a mayor, a municipal council, and various departments that handle specific areas such as urban planning, social services, and public safety. These local governments are tasked with implementing national policies at the local level and addressing the unique needs of their communities with their own policies.

In recent years, municipalities have increasingly adopted AI models to enhance their decision-making processes (Wirtz, Weyerer, and Geyer [140]). AI applications in municipalities can range from optimizing traffic management and improving waste collection to identifying social service needs and planning urban development. AI models help municipal officials analyze large datasets, identify trends, and make data-driven decisions that improve the efficiency and effectiveness of municipal services. However, because municipalities directly impact the lives of citizens, they should critically handle the use

of the models. The European AI Act, proposed by the European Commission in April 2021, aims to establish a comprehensive regulatory framework for artificial intelligence (AI) within the European Union (Commission [21]). The Act seeks to ensure that AI systems used in the EU, and thus Dutch municipalities, are safe, transparent, ethical, and respect fundamental rights. The scope of this actor analysis is on the municipal level. Still, since this AI act is deemed as leading for the further use of AI in municipalities for the coming years, it is deemed as necessary to discuss. The proposal categorizes AI systems into different risk levels and imposes varying regulatory requirements accordingly

- **Unacceptable Risk**  
AI systems that pose a clear threat to safety, livelihoods, and rights are banned. This includes AI applications like government social scoring.
- **High Risk**  
AI systems used in critical areas such as healthcare, education, employment, law enforcement, and essential public services must meet strict requirements. These include rigorous risk assessments, documentation, transparency, and human oversight.
- **Limited Risk**  
AI systems with limited risk are subject to transparency obligations, such as informing users that they are interacting with an AI system.
- **Minimal Risk**  
These systems pose minimal or no risk and are largely exempt from regulation, such as AI used in spam filters or video games.

Given the potential impact on citizens' lives, AI models regarding the human perception of public space would likely be classified as high risk under the AI Act due to the possible impacts on citizens' lives. This is because possible policy advice on the public space could alter the public space, and thus affect how people feel and live in their neighbourhood, as well as possible changes to mobility (car spaces and bike lanes for example). Next, if the results of the analysis of Rotterdam with this model are made public (showcasing which neighbourhoods are more liveable than others), this can affect the housing prices in these areas. Therefore, making the model transparent is of high priority, next to this it also has to be determined by the Municipality whether these results are made public or not. The Municipality of Rotterdam itself has employees who can assess these risks, the possibly high risk is an estimation of the modeller. Additionally, a choice has to be made whether the analysis of these models is made once or used more continuously to monitor areas' liveability. This would alter the risk to be higher if it is used continuously to monitor neighbourhood developments. Municipalities can enhance trust and accountability in their AI systems by following the AI Act. The AI Act's emphasis on transparency ensures that citizens are aware of when and how AI is being used, fostering public trust. For instance, a municipality using AI to manage traffic flow must provide clear information on how the system operates, the data it uses, and the potential impacts on traffic management.

The system consists of municipalities dealing with their citizens, where the municipalities as well as the citizens are themselves influenced by the government. In addition, the citizens in turn affect the municipalities by voting for the municipal council, cooperating with the municipality in civic initiatives, reporting issues to the municipality, and using the municipality's facilities. Citizens also affect the government itself by voting, which in turn determines the budget and direction for municipalities.

Key feedback loops in this system include a decision-feedback loop and a regulation-compliance loop. The decision-feedback loop relates to the way that municipal interventions (like the use of AI models for human perception) impact citizens, whose feedback influences future municipal decisions, and the regulation-compliance loop concerns regulations that guide AI use, like the AI Act, providing feedback to policymakers.

## 5.2. Actor analysis

This research delves into the actors involved in shaping the perception of public space in Rotterdam. Appendix B offers a comprehensive analysis of these actors and their connections to public space per-



ception. Due to the scope of the municipal decision-making, the municipality of Rotterdam itself is not an actor. The municipality however encompasses various departments and entities that directly and indirectly impact public space perception. These actors are responsible for implementing policies and making decisions that shape the urban environment. They also receive financial gains from the taxes of their citizens, as well as financial subsidies from the National Government. Therefore, in this actor analysis, the departments of the municipality of Rotterdam are given as separate actors.

Key actors influencing municipal decision-making regarding public space perception include:

- **National government**  
The national government sets the legislative framework, allocates budgets (so-called municipal funds (Rijksoverheid [112])), and influences policies that affect municipalities like Rotterdam. This includes funding for infrastructure projects and social services, which are crucial for shaping public spaces (Rijksoverheid [113]).
- **European Commission**  
The European Commission plays a role in creating regulations impacting AI use for European municipalities. This includes directives on proposed legislation on artificial intelligence (AI Act), which set standards for AI use in public administration and safeguard citizen rights (Commission [21]).
- **Citizens of Rotterdam**  
The residents of Rotterdam are directly affected by municipal decisions and policies regarding public spaces. They provide feedback through civic engagement, participate in local initiatives, and influence decision-making processes through voting and community involvement.
- **CROW**  
CROW is a measurement method used to assess the objective quality of public spaces. This method evaluates the condition of public facilities, thereby determining how the public space is evaluated objectively.
- **Department of Stedelijk Beheer**  
Stedelijk Beheer is concerned with creating a vision of how to manage public spaces. They determine the operations of public space management and ensure that these areas are clean, safe, and functional. Their management style has recently shifted to a value-driven approach, emphasizing inclusiveness, sustainability, and social aspects.
- **Department of Stedelijke Inrichting**  
This department focuses on urban planning and development in Rotterdam. They guide the spatial layout, architectural design, and economic aspects of new developments, influencing the physical and visual characteristics of public spaces. Their decisions impact urban aesthetics, accessibility, and the overall quality of life for residents. They contribute to a safe, physical, and inclusive living environment, ensuring an attractive and clean public space.
- **AMOR**  
Asset Management Openbare Ruimte (AMOR) manages all assets in Rotterdam, ensuring they function well and look well-maintained. This contributes to the functionality and aesthetics of public spaces.
- **Onderzoek en Business Intelligence (OBI)**  
OBI plays a crucial role in providing data-driven insights and intelligence to the Municipality of Rotterdam. They conduct research, analyze data, and provide valuable information on various topics such as economy, health, education, and urban development. OBI's work enhances the municipality's ability to make informed decisions that impact public space perception and citizen well-being.
- **Advanced Analytics department**  
A department from OBI has the technical knowledge to work with complex data analyses, machine learning models and computer vision models. For technical models, this department is responsible for the creation and management of these models. They work mostly on a project

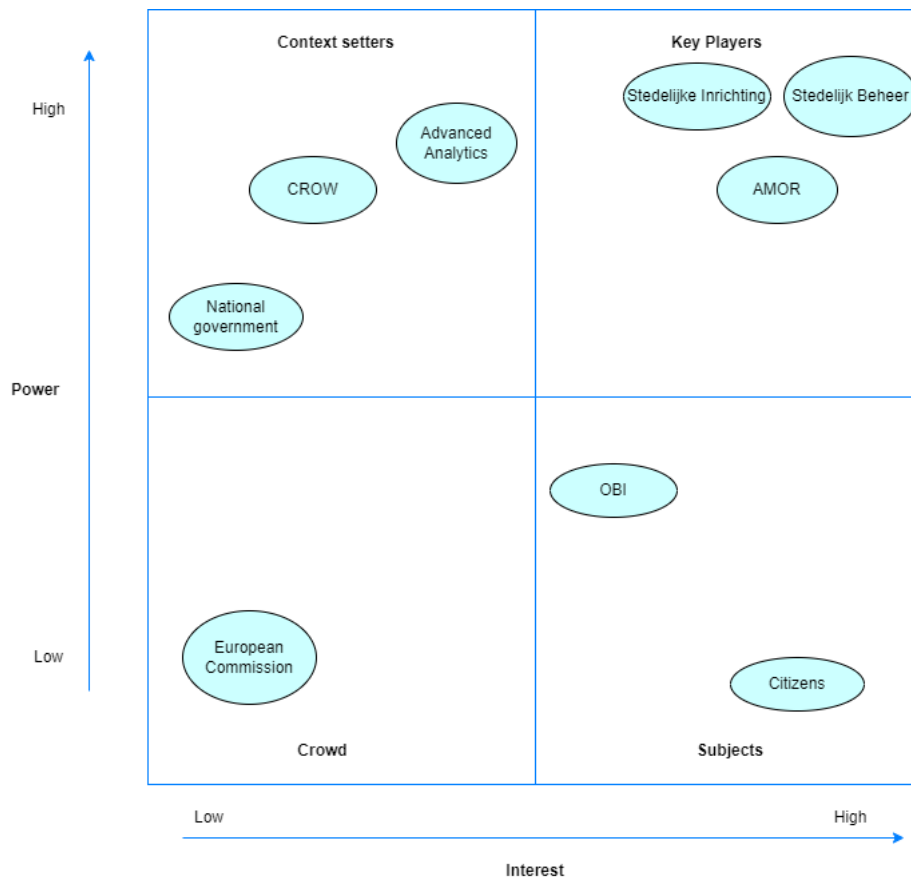
basis, therefore being well suited to work on specific projects like CVDCMs for the perception of the public space.

### 5.2.1. Power-Interest Matrix

The Power-Interest Matrix (Mitchell, Agle, and Wood [94]) categorizes actors based on their influence (power) and engagement (interest) in municipal decision-making processes regarding public space perception. This tool helps identify stakeholders' roles and their impact on shaping urban policies and environments. Power refers to the ability or capacity of stakeholders to influence the decision-making process. Interest refers to the degree to which stakeholders are affected by or concerned about the outcomes of decisions related to public space.

- High Power, High Interest (key players)  
These actors have significant influence and are deeply involved in decision-making processes. They actively shape policies and regulations that impact public space perception.
- High Power, Low Interest (context setters)  
Actors in this category may have considerable authority over municipalities but may not always be directly engaged in specific urban development decisions.
- Low Power, High Interest (subjects)  
These actors have a high interest in municipal decisions affecting public spaces but may not individually hold significant power.
- Low Power, Low Interest (crowd)  
Actors in this category have minimal influence and engagement in municipal decision-making processes regarding public spaces.

By analyzing the roles and interactions of these actors, municipalities like Rotterdam can better navigate complexities, foster collaboration, and ensure that policies and projects effectively enhance public space perception and citizen well-being. Figure 5.1 shows the Power-Interest Matrix.



**Figure 5.1:** Power-Interest Matrix of the actor analysis. Power refers to the ability or capacity of stakeholders to influence the decision-making process. Interest refers to the degree to which stakeholders are affected by or concerned about the outcomes of decisions related to public space.

### High Power, High Interest

The Department of Stedelijke Inrichting, responsible for urban planning and development, wields power through the implementation of development projects that affect the city's layout and functionality. They guide the spatial layout, architectural design, and economic aspects of new developments, influencing the physical and visual characteristics of public spaces. Stadsontwikkeling's decisions impact urban aesthetics, accessibility, and the overall quality of life for residents. Their interest lies in long-term strategic planning, focusing on how urban spaces evolve and improve over time. Therefore, their interest is very high, and so is their power. The way building is done in cities like Rotterdam is that orders come in from the municipality but also from contractors, to build houses or other places in the outdoor space. The Stedelijke Inrichting Department then sits down with these parties to come up with a plan of action. The municipality can then include their values such as sustainability and the Rotterdam Style, in order to steer the building plans to what the municipality wants to see. From the contractor's point of view, the values of sustainability and Rotterdam Style are not often included in the designs, let alone the liveability of the public space. Therefore, the knowledge that comes from these models can be used in the design of all building plans of the municipality, in order to focus even more on liveability-oriented building of the outdoor space. Here, the interest of Stedelijke Inrichting is thus very high, and thus their strength is even higher because they can ultimately shape the plans towards building with liveability as a goal.

The Asset Management Openbare Ruimte (AMOR) department manages all assets in Rotterdam, ensuring they function well and look well-maintained. This contributes to the functionality and aesthetics of public spaces. They do the daily operational side needed to maintain the public space. They operate on the vision of Stedelijk Beheer and therefore have less power than that department, but still have high interest due to their correct connection on how the public space is designed and maintained, as well as

their asset management style of the public space being influential on the aesthetic of the public space.

The department of Stedelijk Beheer is concerned with creating a vision of how to manage the public space. They set how the public space is managed, and determine thus how the operations of public space management are done. They initiate the cleanliness, safety, and functionality of public spaces, which directly impacts how these areas are perceived by the public. In recent years, there has been a shift in the management approach of Stadsbeheer, moving from a functional management style focused on the objective state of outdoor objects to a value-driven management style. This new approach emphasizes inclusiveness, sustainability, and social aspects, for example, aligning more closely with subjective human perceptions of outdoor space. Value-driven management prioritizes the holistic view of human experience, considering how the maintenance and appearance of public spaces contribute to overall well-being and satisfaction. This shift towards value-driven management is particularly relevant to this research, as it underscores the importance of subjective and holistic human perception in the planning and maintenance of public spaces. By focusing on how individuals perceive and interact with their environment, Stadsbeheer aims to create spaces that are not only functional but also enriching and supportive of community values. Therefore, their interest is high, as well as their power.

#### High Power, Low Interest

The CROW system is used as a measurement method to assess the objective quality of public spaces. This method is applied to evaluate the condition of all public facilities, thereby determining how the public space is evaluated objectively. This provides them with high power, as they determine this objective quality of the public space which is used by the municipality for management of the public space. Their interest is low, as they only set a measurement, not operate in the public space.

The National Government provides budget allocations and legislative frameworks that significantly impact local policies. While they hold considerable power, their direct involvement in the specifics of public space perception in Rotterdam is less pronounced. Their primary interest lies in achieving national objectives, with a moderate focus on the local public space experience. The government has a lot of power because of their financial strength, but their focus is not on Rotterdam's outdoor space, so their interest is low.

Advanced analytics provides the technical knowledge for further analyses at the Municipality. Due to their knowledge of data analysis and model creation, more complicated technical questions are tackled by them, in comparison to OBI, which researches the less technical questions. In terms of perception of the public space, they are the departments that can manage and use computer vision models to monitor the public space, such as CVDCMs. This makes their power high as they are the only department capable of creating and managing these models, while their interest is medium as they will have to be directly involved in the perception of the public space when they create technical solutions for this, but are still only focused on the creation of these models, and not in the public space perception.

#### Low Power, High Interest

Onderzoek en Business Intelligence (OBI) provides critical data and insights that inform policymaking in Rotterdam. Their power is more indirect and thus lower, influencing decisions through the provision of data, as they only can act on research requests by departments like Stadsbeheer. Their interest is high due to their essential role in shaping urban planning and policy decisions. Their research helps the municipality understand and address public space perception issues effectively. Their research can have a lot of impact on how outdoor space is created by urban development and management, their power is medium but their interest is not only in analysing citizens' perceptions of outdoor space, so their interest is somewhat lower.

The Citizens of Rotterdam are highly interested in the quality and usability of public spaces. As end users, their daily lives are directly affected by the state of these areas. They influence public space perception through feedback, participation in local initiatives, and the use of reporting tools like the MELDR app. Their involvement is crucial in shaping public spaces to meet community needs and preferences. The citizens of Rotterdam are directly affected by all public space adaptations, or the lack of adaptations, so their interest is very high, only their power is low.

#### Low Power, Low Interest

The European Commission holds regulatory power over data protection and the ethical use of AI, which impacts how human perception studies are conducted and implemented in public space analysis. While their influence is significant in ensuring compliance with regulations like the AI Act, their direct involvement in the specifics of public space perception is limited. Their primary interest lies in maintaining ethical standards and transparency in AI applications. The European Commission currently has little power as there are no hard restrictions on AI models for modelling public space, the AI Act does not restrict municipalities to certain actions. They are in this actor analysis because they are therefore influential in how public space is researched in this research, but their interest is in many other things than citizens' perceptions of public space.

### 5.3. Interactions and Dependencies

Understanding the interactions and dependencies between these actors is critical to comprehending how public space perception is shaped in Rotterdam. The municipality, including its various departments, relies on the national government for legislative frameworks and budget allocations. In turn, the national government depends on municipalities to implement policies and achieve national objectives. The municipality must also comply with regulations set by the European Commission, such as the AI Act guidelines, ensuring ethical and transparent AI use. Citizens provide feedback and participate in local initiatives, influencing municipal decisions, while the municipality depends on this engagement to shape public spaces that meet community needs.

OBI provides data-driven insights that inform Stedelijk Beheer's and Stedelijke Inrichting's operational decisions. Stedelijk Beheer's management and maintenance activities generate data that OBI analyzes, creating a feedback loop for continuous improvement. OBI's policy advice directly influences Stedelijke Inrichting, which in turn necessitates further research by OBI. When this research involves highly technical aspects like machine learning models, Advanced Analytics collaborates with OBI.

Stedelijke Inrichting's planning decisions directly impact the areas managed by Stedelijk Beheer. Effective collaboration ensures that urban development projects align with maintenance and operational capabilities.

### 5.4. Expert Analysis

Conducting an expert analysis is essential for effectively evaluating and interpreting the results of visual XAI techniques used in CVDCMs. By performing the actor analysis and examining the dependencies and interactions between various stakeholders, we can identify key actors capable of working with models like the CVDCM within the municipality. This identification is crucial for ensuring that the model's outputs are both relevant and actionable. Expert analysis techniques for XAI provide a structured approach to further analyze these results.

According to Doshi-Velez and Kim [27], an effective evaluation framework for the interpretability and explainability of model results includes two relevant methods: application-grounded evaluation and human-grounded evaluation. Application-grounded evaluation involves domain experts conducting human experiments within the context of a real application. This method directly aligns with the system's intended objective but can be costly due to the need for multiple domain experts to dedicate time to provide critical judgment on the results. On the other hand, human-grounded evaluation involves simpler experiments with non-domain experts to gauge the explanation's quality, which is less expensive but might not fully capture the effectiveness of the explanation.

At the municipality, experts on human perception of public space are widely available, making application-grounded evaluation well-suited for Rotterdam. This is crucial because these experts can validate whether the results from visual XAI techniques align with their findings in urban planning and public space management. Furthermore, one could argue that citizens themselves are experts in the perception of public space, given their daily interactions with it. The municipality's direct contact with the citizens of Rotterdam allows for incorporating their feedback into the evaluation process, potentially broadening the scope of the application-grounded evaluation and providing a more comprehensive

view.

Integrating expert analysis within this stakeholder framework underscores its importance. A thorough stakeholder analysis helps identify key actors and their interactions, revealing the dependencies that impact decision-making processes. By ensuring that expert evaluations are part of this framework, the research not only benefits from technical validation but also gains practical insights that can enhance the usability and applicability of XAI techniques in urban planning. This holistic approach is essential for developing meaningful XAI applications that can effectively inform policy decisions and improve public space management.

## 5.5. Conclusion

With the system analysis and actor analysis, it becomes clear how Computer Vision Enriched discrete choice models with LIME explanations can be used effectively within the Municipality of Rotterdam. These models should be handled by actors who have combined technical and policy expertise, particularly concerning public space perception (so an actor does not need to have both the technical and policy knowledge, but the group of actors need to have it in combination). The Onderzoek en Business Intelligence (OBI) department Research Team together with the Advanced Analytics department is identified as the ideal group to implement these models, leveraging their technical knowledge to provide valuable input for policy advice. Advanced Analytics has the technical knowledge needed for computer vision models, as well as eXplainable AI, and can work on a project basis for this specific model, while OBI has the teams to conduct research and prepare this for the right departments. Therefore, together they can form the most suited group to analyze and work with CVDCMs. Collaboration with Stedelijk Beheer and AMOR in the expert analysis is essential to incorporate practical insights about public spaces, thus creating comprehensive recommendations for both Stadsbeheer and Stadsonwikkeling. This creates a square workflow between Advanced Analytics, OBI and the expert panel from Stedelijk Beheer and AMOR to provide the most well-suited insights from these models for the municipality. This coordinated approach ensures that the use of AI models is both technically sound and aligned with the needs of the urban environment, ultimately enhancing the quality and usability of public spaces in Rotterdam, as well as complying with the AI Act. Additionally, with these possible new insights gained from the model's outcomes, Stedelijk Beheer can develop the urban public space with more consideration on liveability thereby creating better cities for all citizens. In cooperation with the contractors who build the public space and the houses, the importance of creating a liveable public space becomes more and more evident.

# 6

## Face validity analysis of Rotterdam's utility-based liveability score

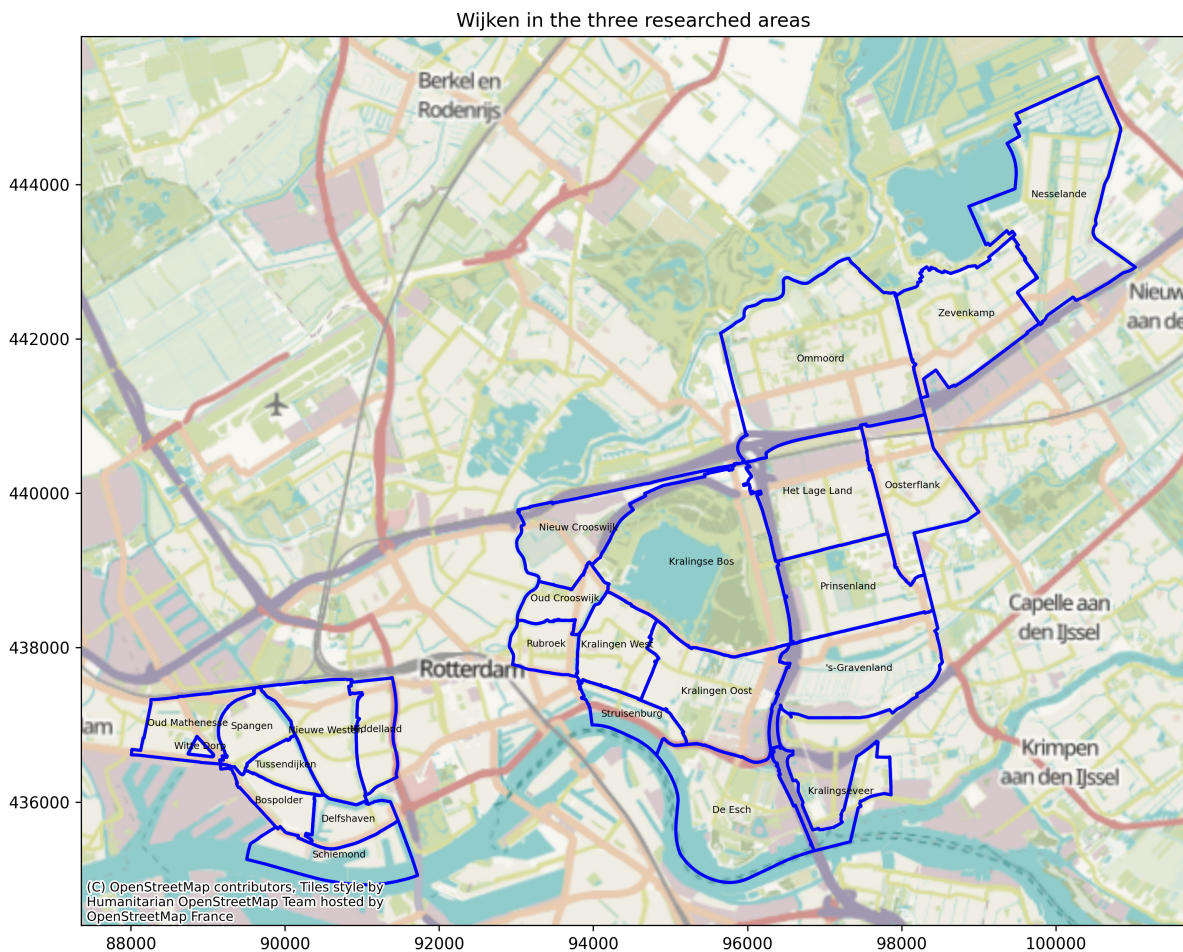
The methodology and implementation chapter establishes the foundation for applying the CVDCM to Rotterdam, aiming to evaluate the model's outcomes. This chapter provides a practical assessment of the CVDCM within the municipal context through a face validity analysis of the Rotterdam research area. Each neighbourhood in the research is discussed, and this discussion showcases one practical usage for this model in the municipality of Rotterdam, as discussed in the previous chapter. The face validity analysis builds on the methodology and implementation by evaluating the model's decision behaviour, which is crucial for determining whether LIME enhances the understanding of the CVDCM's decision-making processes compared to a standard face validity analysis. To effectively measure the impact of XAI on the interpretability of the CVDCM, it is essential to compare the interpretability before and after the implementation of XAI. Thus, a preliminary analysis of the model's behaviour through face validity is necessary. Additionally, the knowledge about liveability gained from the literature review is used to inform the face validity analysis, ensuring that the evaluation of the model's decision behaviour is grounded in established theoretical insights about urban liveability. Although the theoretical basis from the methodology indicates that parameters  $\beta_k$  and  $w$  from equation 3.4 reflect decisions related to the utility-based liveability score, they lack behavioural significance. This section seeks to answer the fundamental question: what insights has the CVDCM learned about street-level conditions?

### 6.1. Research location

The research location choice is based on a few criteria. First, the location needs to have a varied street view collection. The research area thus needs to encompass enough different street views, such that every type of street view is included in the analysis. This requires the area to have different building styles, the presence of nature and water, and also of more and less densely populated areas. A second requirement is that the subjective evaluation of the perceived subjective liveability in the neighbourhood needs to differ from the perceived objective liveability. This allows for an analysis which is useful for the municipality of Rotterdam, as it also allows them to analyze possibly why there is a difference in the objective and subjective perceived liveability.

The "Wijkprofiel" (Rotterdam [115]) of the municipality of Rotterdam refers to a detailed description of various aspects of each neighbourhood within the city on how well they perform on safety, the physical environment and social, measured subjectively and objectively. It Wijkprofiel provides information for determining interesting differences between neighbourhoods for the research area. Based on 'de Wijkprofiel' and discussions with the Municipality of Rotterdam about the criteria mentioned, three areas were chosen as research locations. The areas are Prins Alexander, Kralingen-Crooswijk and Delftshaven. Prins Alexander is a more recently built area of Rotterdam, thus containing more newly developed style buildings. The other two were built earlier, pre-second world war. Kralingen contains more nature, such as the Kralingse Bos, and Delftshaven contains a historic sight and is more densely

populated. Kralingen has a lot of nature, but also densely built areas with students and social housing. Delfshaven lacks a high amount of nature and is more densely built in streets, containing vibrant streets and social housing. and a high mobility network for vehicles. Therefore, they represent a vibrant mix of build styles, build years, green spaces, and demographic differences. In these three areas, the subjective (which is measured via surveys in the neighbourhood) and objectively perceived liveability (which is measured by scores of the municipality) differ, meeting the first criteria. Subjectively perceived liveability is valued lower than objectively perceived liveability in all areas. Therefore, the research area is the neighbourhoods of Delftshaven, Prins-Alexander and Kralingen-Crooswijk. These areas with their neighbourhoods are shown in Figure 6.1.

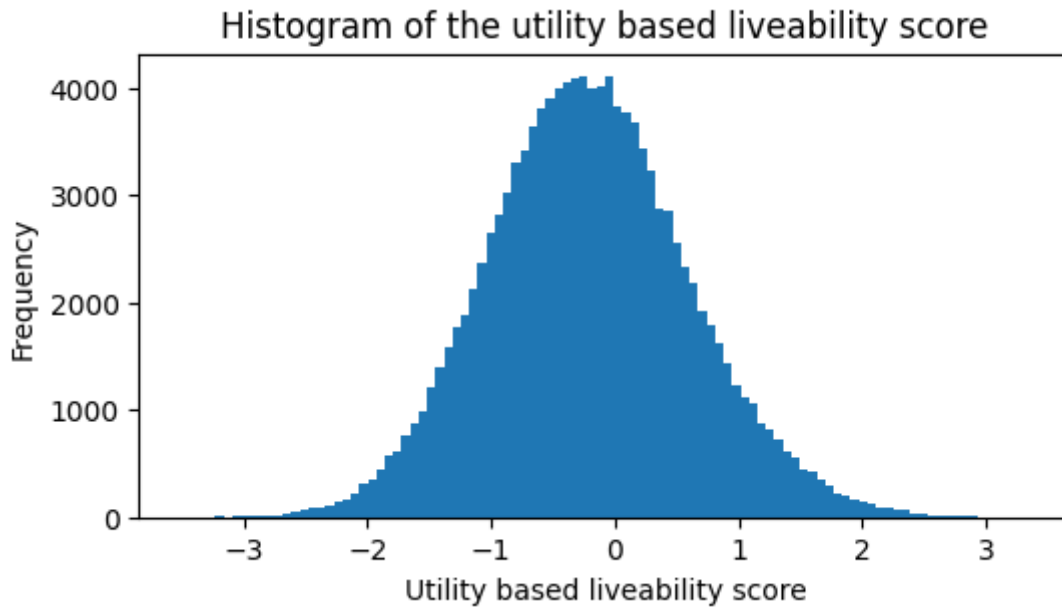


**Figure 6.1:** All neighbourhoods of the research area's geographical location

Each area contains multiple neighbourhoods. Delftshaven contains the following neighbourhoods: Bospolder, Delfshaven, Middelland, Nieuwe Westen, Oud Mathenesse, Schiemond, Spangen, Tussendijken. Prins-Alexander contains the following 'Wijken': Het Lage Land, Kralingseveer, Nesselande, Ommoord, Oosterflank, Prinsenland, s-Gravenland, Zevenkamp. Kralingen-Crooswijk contains the following 'Wijken': de Esch, Kralingen Oost/Bos, Kralingen-West, Nieuw Crooswijk, Oud Crooswijk, Rubroek, Struisenburg.

The image data retrieval for these three areas, with the front and back view of the homes, resulted in around 120000 images. For each image, the utility-based liveability score is calculated. Figure 6.2 shows the distribution of the utility-based liveability scores, and Table 6.1 shows the overall results of this analysis.





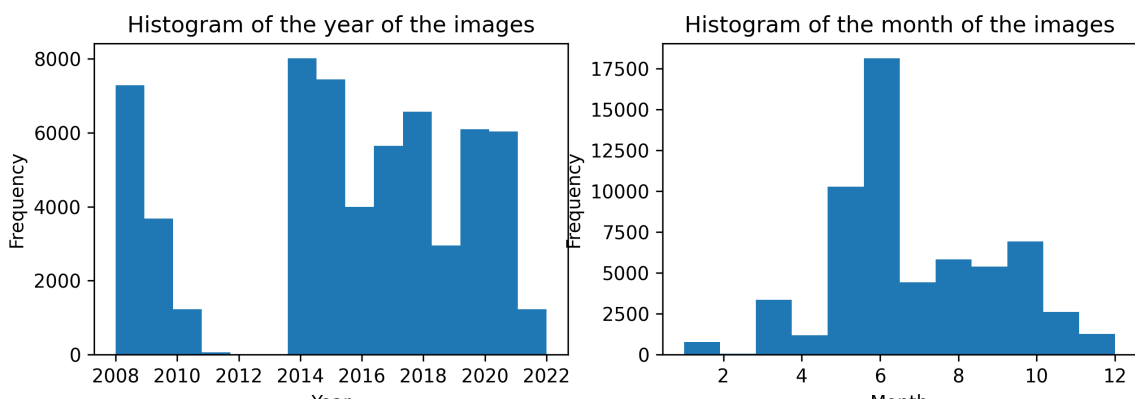
**Figure 6.2:** Histogram of the utility-based liveability score of the street view images from the research location

Figure 6.2 illustrates that the distribution of scores skews to the left. Table 6.1 indicates that the mean of these scores is around -0.2, consistent with the observed leftward skewness in the histogram. While the score can theoretically range from negative infinity to positive infinity, in this context, utility-based liveability scores are confined to the interval between -4 and 4. It's important to note that a negative score does not inherently denote poor liveability. Rather, it signifies that negative influences within the street view outweigh positive ones, or that specific aspects exert a disproportionately negative impact, overshadowing any positive elements. Theoretically, a completely blank street view would garner a score of 0, presuming the absence of both the negative and positive elements.

Count	Mean	Std	Min	Max
120519	-0.212273	0.816288	-3.502888	3.341217

**Table 6.1:** Utility-based liveability score statistics

Figure 6.3 shows the distribution of the month and year of when the street view images in this research are taken.



**Figure 6.3:** Distribution of the utility score: Left, the year it is taken in and right the month it is taken in

As can be seen, most pictures were dated from 2010, 2014, and 2015. The month where it was taken was mostly in the 7th month, June. As stated by (Nasar [102], Cassidy [14]), the weather and the amount of sunlight can affect how a person experiences the public space. Summer months, like the 7th month (July), are on average more sunny. Figure 6.4 shows the distribution of average utility scores per month in our dataset, where February, April, and December score lower.

The analysis reveals that the majority of pictures were dated from 2010, 2014, and 2015. Moreover, the most popular months of capture were June and July, both summer months and tend to be characterized by higher average levels of sunlight. In studies such as (Nasar [102], Cassidy [14]), it is suggested that weather conditions, including sunlight exposure, can significantly influence individuals' experiences of public spaces, and thus the utility-based liveability score. Therefore, a higher utility score should be expected in these months. Figure 6.4 depicts the distribution of average utility scores per month within our dataset. Interestingly, lower scores are noted for February, April, and December, while the summer months score around the average on the utility-based liveability score.

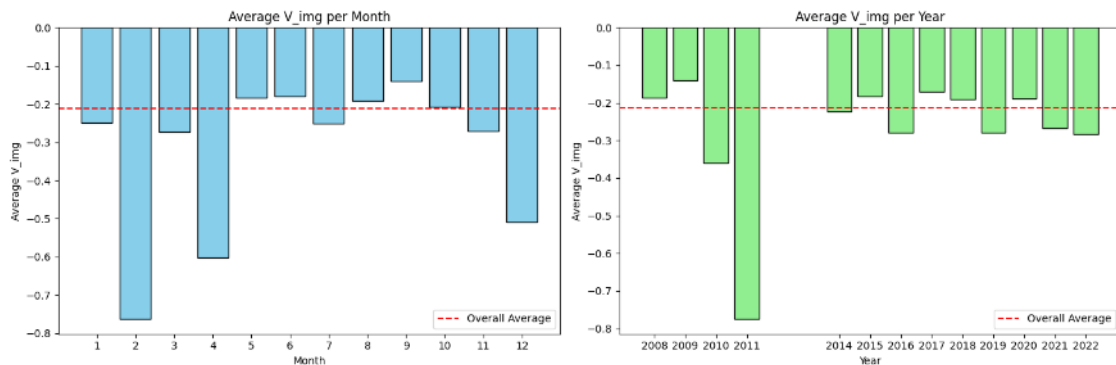


Figure 6.4: Distribution of the average utility score per month and year

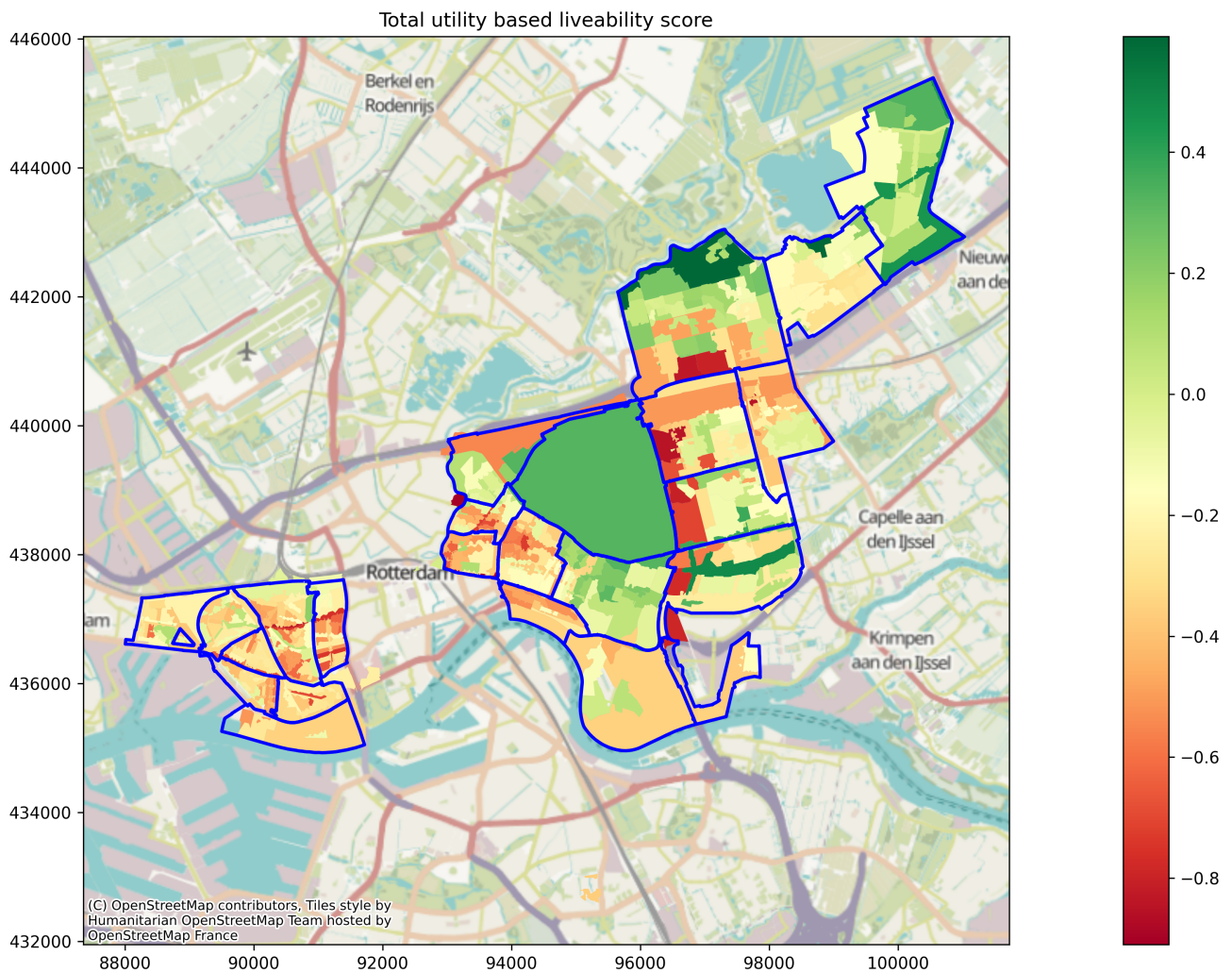
The observed variation in utility scores across different months is indeed notable. However, it's important to note that no correction needs to be applied for this month-based differences. In the model used for this analysis, the utility score is derived from the combined influence of the image's feature map and the estimated month of the year. As such, it's challenging to isolate the specific effect of the month, given the integration with the feature map in generating the overall score. For instance, the feature map may detect a tree in a street view taken in December without leaves, different from the representation in summer months, thus still recognising the existence of a tree, if deemed influential. This feature map usage thus already deals with the monthly differences, and no after-correction is needed. Consequently, the effects of the month are intertwined with the image's features in the scoring process.

## 6.2. Utility-based liveability scores of selected areas

In the Netherlands, zip codes are structured into hierarchical levels, each offering varying degrees of pinpointing geographical locations. At the lowest level is the level 6 zip code, or PC6, which serves as the most precise identifier and is used for addressing mail. An example of a PC6 would be 3023DD, denoting a specific zip code in Rotterdam typically corresponding to a single street or cluster of addresses. Moving up one level is the zip code zone, or PC5, which delineates smaller areas compared to PC6, such as neighbourhoods or groups of streets. For instance, 3023D represents a distinct area within Rotterdam. At the highest level is the zip code area, or PC4, which encompasses entire neighbourhoods, villages, or cities. An example would be 3023, signifying a specific neighbourhood within Rotterdam. In summary, each level of Dutch postcodes offers varying levels of detail and precision in identifying locations in the Netherlands. PC6 provides the most detail, followed by PC5 and then PC4. For this study, PC4 and PC5 levels are used. PC5 level focuses on these specific zip codes in the neighbourhoods already discussed. PC4 is the zip code of these specific neighbourhoods. For a more detailed view of the results, PC5 will be used, and for a more general and average view, PC4 will be used.

Not all level 5 zip codes neatly align with the corresponding level 4 areas, as evidenced by discrepancies observed in Kralingen Crooswijk and Prins Alexander. This inconsistency arises due to the limitations of working with level 5 zip code areas, but for this analysis, it is considered negligible. Therefore, PC5 will be used as the primary reference for mapping and analysis together with PC4. In Appendix E, a combined mapping of these two areas can be found, serving as a reference for further analysis to understand how specific zip codes may fall within different PC4 areas.

Figure 6.5 illustrates the outcomes of the modelling process regarding the utility-based liveability score of the designated areas at the PC5 level in Rotterdam. The blue outlines show the neighbourhoods within the area, which are the PC4 level areas as discussed.



**Figure 6.5:** Average utility-based liveability score per zip code level 5 in the research area

Here, the spatial distribution of the utility-based liveability scores across the map of Rotterdam is shown. The map represents the average utility-based liveability score per PC5 zipcode. During the modelling process, each street view is assigned a geographical point on the map corresponding to their score. This mapping procedure is detailed in Appendix C, where all geographical points are plotted on the map. At the PC5 level zip code, the utility score is averaged of all these points in the zip code areas to provide a clearer visualization of the utility score mapping.

Upon examination, it is evident that the Delftshaven area exhibits relatively average lower scores, whereas the Prins Alexander area encompasses zip codes with the highest scores, alongside some of the lowest-scoring zip code level neighbourhoods. Analyzing the characteristics of the best and worst-

scoring neighbourhoods enables the modeller to identify elements within street views that contribute positively and negatively to the utility-based liveability score.

Subsequently, the 10 highest- and worst-scoring zip code level neighbourhoods, along with their respective neighbourhood are shown in table 6.3 and 6.2.

Zipcode	Average utility score	Buurt name
3036J	-0.909536	Oud Crooswijk
3031E	-0.899410	Rubroek
3067K	-0.847769	Oosterflank
3068N	-0.828203	Ommoord
3022A	-0.822582	Nieuwe Westen
3066P	-0.805066	Prinsenland
3068C	-0.803068	Ommoord
3065W	-0.788048	Kralingseveer
3065D	-0.772079	Kralingen Oost
3021B	-0.762565	Middelland

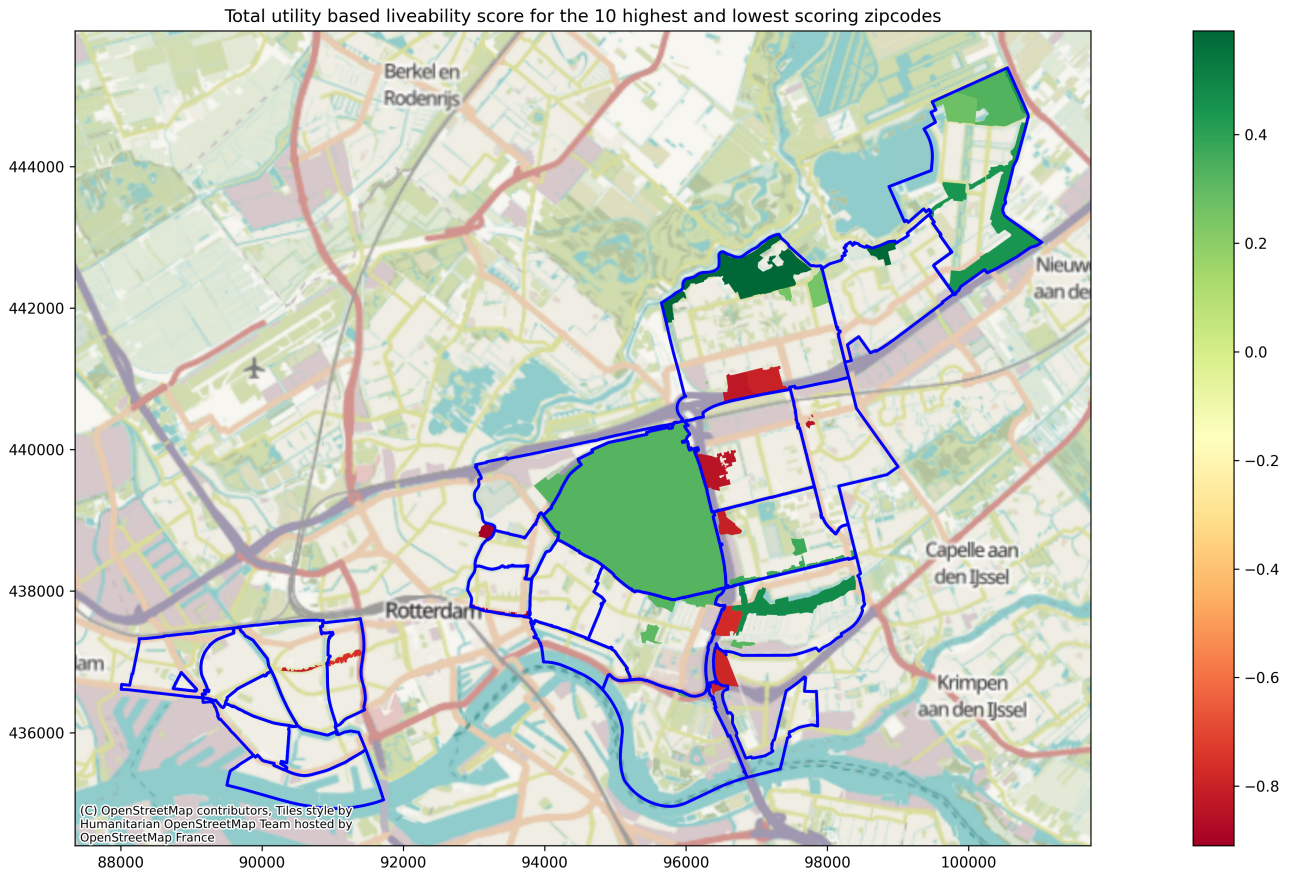
**Table 6.2:** The 10 lowest average scoring zip codes (level 5)

It can be seen that the worst performing zip code level neighbourhoods are not centred at one or two specific PC4 neighbourhoods, but are more scattered around the research area. The highest scoring PC5 neighbourhoods are shown in table 6.3.

Zipcode PC5	Average utility score	PC4 neighbourhood
3069L	0.590989	Ommoord
3065A	0.483151	's-Gravenland
3059L	0.445662	Nesselande
3066R	0.370016	Prinsenland
3059S	0.331064	Nesselande
3062C	0.326767	De Esch
3065C	0.317647	's-Gravenland
3062R	0.309474	Kralingen Oost
3059Z	0.271048	Nesselande
3069A	0.256409	Ommoord

**Table 6.3:** The 10 highest average scoring zip codes (level 5)

This figure already shows that in Nesselande, multiple PC5 neighbourhoods with high scores are present. Figure 6.6 further shows the locations of the PC5 zip code areas with the highest and lowest average utility-based liveability scores. It can be seen that multiple areas in Nesselande score well, just as the Kralingse Bos is scoring well, also Kralingen Oost. Kralingseveer, het Lage Land and Middelland are showing lower scores.



**Figure 6.6:** Locations of the 10 highest and lowest scoring zip code areas

These PC4 neighbourhoods which perform relatively good and bad are taken for further inspection of their street view utility-based liveability scores.

Before heading into the examination of the specific neighbourhoods discussed earlier, it is beneficial to conduct a preliminary evaluation of the model's overarching performance. To generate a diverse array of images with disparate scoring profiles, the analysis employs quantile plots. Quantiles, in statistical explanation, divide a dataset into intervals featuring an equivalent number of data points. Thus, creating six quantiles entails assessing six intervals, each comprising an even number of data points, i.e., street-view images. A quantile plot, therefore, acts as a grid of street view images, wherein each row signifies a quantile alongside the corresponding range of utility scores. The quantity of images per row is arbitrarily determined by the analyst to afford a comprehensive spectrum of images, creating a comprehension of the model's scoring patterns. Figures 6.7 and the two figures in Appendix F.1 and F.2 show a quantile-based scrutiny of image scores.



**Figure 6.7:** Quantile plot of the image utility-based liveability scores

In Figure 6.7, it is clear that street views featuring highways receive notably lower scores, as evidenced by quantile 1. Conversely, scenes characterized by abundant greenery, trees, and bodies of water yield higher utility scores, as observed in quantile 6. The distinctions become less pronounced in the intermediary quantiles, necessitating further examination. Figure F.1 also shows lower scoring of highway roads and other industrialized places, shown by the presence of the colour grey in the street views. Notably, in quantile 2, a highway scene with extensive tree coverage likely contributes to a higher score than other highway roads in images. Similarly, images with conspicuous greenery or water sources continue to correlate with higher utility scores across quantiles 4, 5, and 6. However, nuanced differences are present as well; quantile 4 shows scenes featuring both pavement and roads, whereas quantile 5 showcases a green landscape with fewer trees but lush ground cover. In contrast, quantile 6 presents scenes either abundant in green leaves or the absence of paved surfaces. The fifth quantile includes a highway scene with high greenery and residential structures, deviating from the prevailing trend of low-scoring highway views.

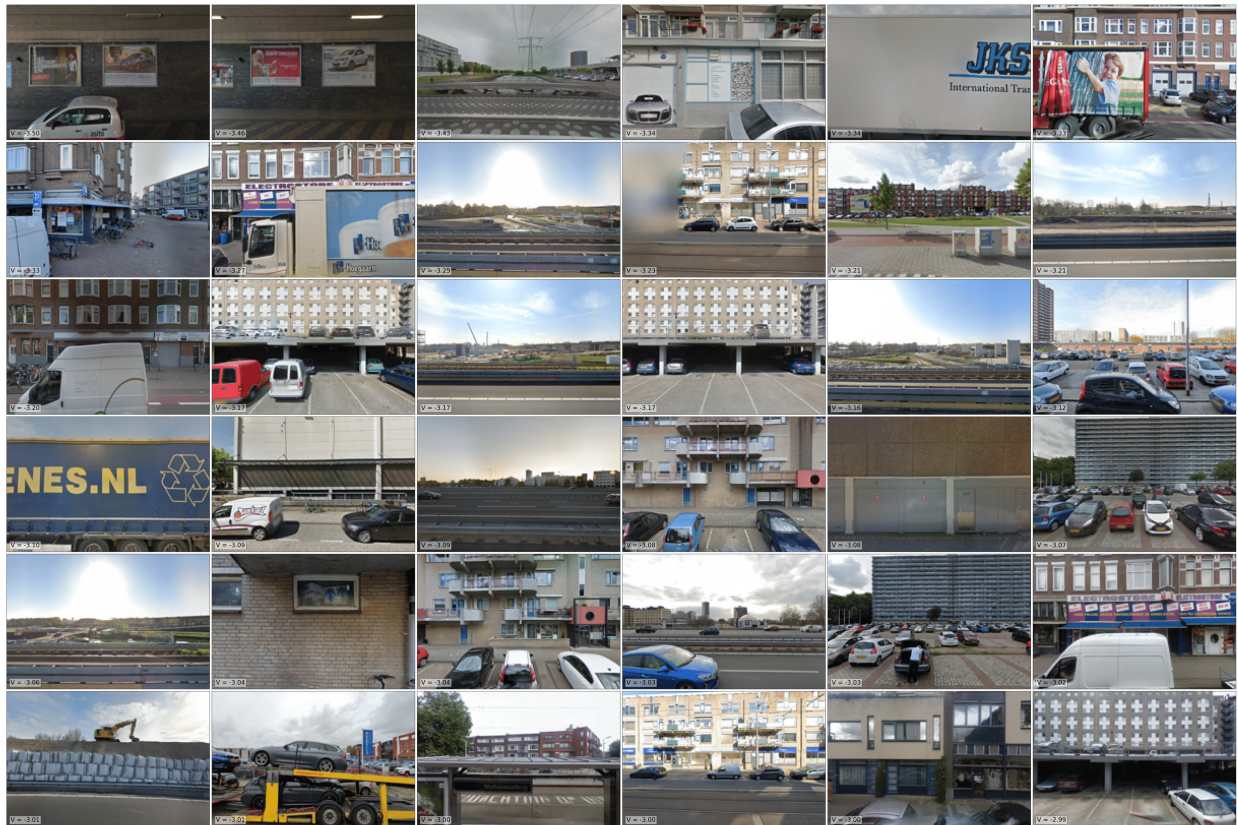
Figure F.2 demonstrates similar patterns. Notably, highways consistently rank poorly in the first quantile, alongside a bicycle lane with a bridge, suggesting a negative impact on utility scores by nearby concrete bridges. Conversely, scenes featuring natural elements continue to have higher ratings. Notably, the second and fourth images in quantile 6 depict conventional residences with limited greenery yet score comparably well. Upon closer examination, it becomes apparent that homes with prominently visible brown exterior bricks tend to score higher, consistent with observations in quantile 5. These building style choices could be influencing the higher utility scores. Furthermore, residences depicted in quantile 6 appear brighter in comparison to those in quantile 2. Remarkably, the fourth image in quantile 6, depicting shaded homes with minimal greenery, stands out due to the absence of vehicular infrastructure, thereby leaving open free space for residents. Notably, the highway segments present across multiple quantiles exhibit variations in green density, suggesting that the amount of green leaves

on trees or bushes may exert an influence on the model's scoring algorithm. In quantile plots, broader trends in the model's decision-making process can be analyzed and discussed. Notably, quantiles 1 and 6 represent extremes in scoring, thus offering clear insights into the most influential aspects of the model's decision-making. To deepen the analysis of these high and low scores, detailed descriptions of the highest and lowest-scoring images are presented in the subsequent figures. This provides a more detailed exploration of the factors significantly influencing the model's assessments. Figure 6.8 and Figure 6.9 respectively showcase the street views with the highest and lowest utility scores.



**Figure 6.8:** The highest scoring street view images

The findings regarding the street views with the highest utility scores underscore the pronounced significance of greenery. The model consistently assigns higher utility value to vegetation, emphasizing the pivotal role of lots of greenery in boosting utility scores. Additionally, the presence of water emerges as a contributing factor to higher scores. Moreover, plenty of open space, often characterized by the absence of densely clustered residential structures, is a recurrent feature in scenes having high utility scores. Notably, an absence of vehicular traffic is evident, although it is interesting to note that the highest-scoring image includes a vehicle. While this vehicle may not be readily interpretable to the model due to the limited visibility (only the top portion of the car is visible, and the car's colour is black), it presents an interesting image. It is plausible that the model interprets the dark colour and reflective properties of the car's surface as indicative of water, thus attributing a higher score. However, this hypothesis remains speculative and needs further investigation. Additionally, the absence of roads is prominently observable in these high-scoring scenes, further highlighting the preference for unobstructed views without vehicular infrastructure.



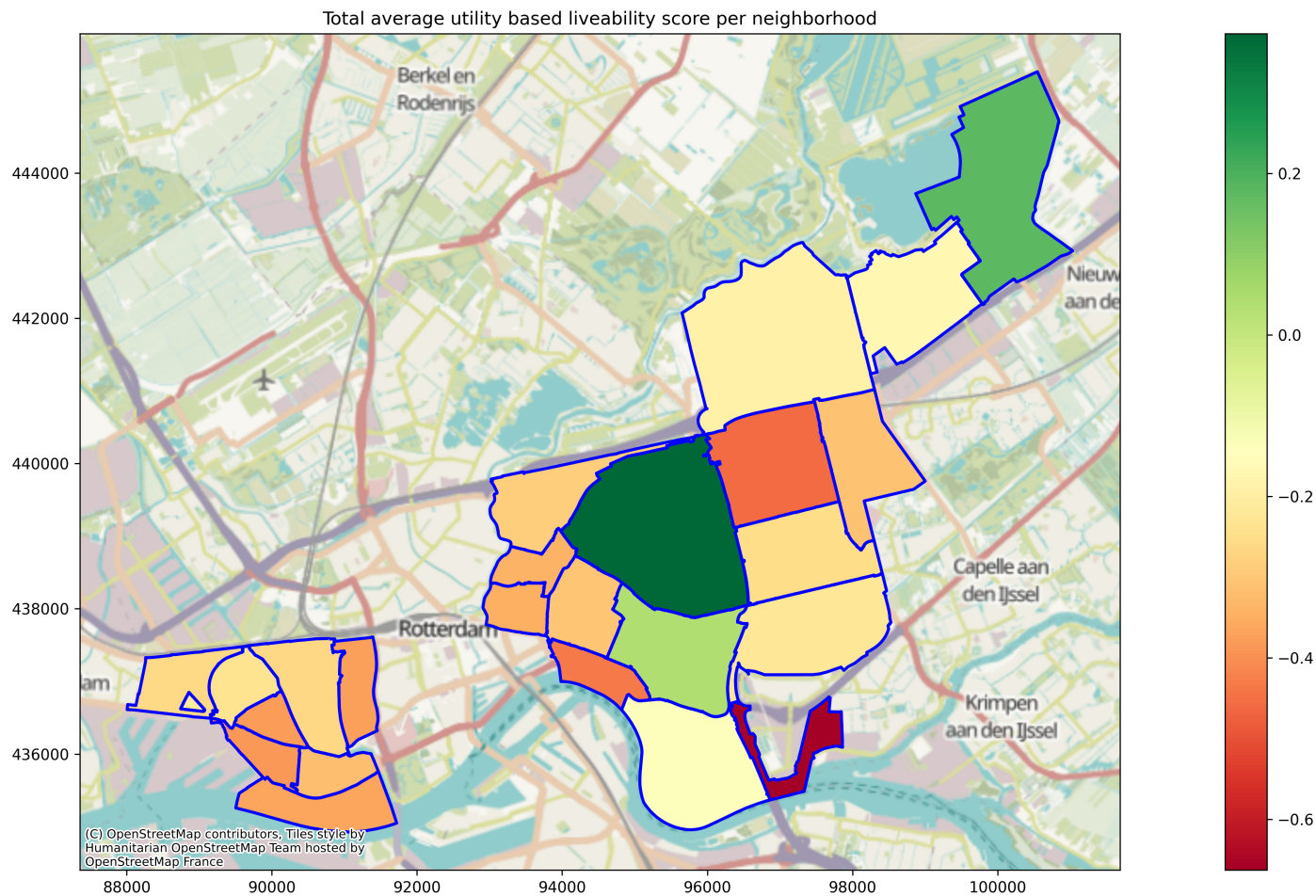
**Figure 6.9:** The lowest scoring street view images

For the lowest utility scores, as depicted in Figure 6.9, the patterns are also quite straightforward and comprehensible. The presence of highway roads and cars both contribute negatively to the utility score. Additionally, the prevalence of the colour grey is evident across all images.

The examination of both the highest and lowest scores provided by the model thus far has shed light on the model's possible behaviour and helped identify factors possibly determining the utility-based liveability score. It is evident that greenery, particularly trees, along with the presence of water, a lack of cars and roads, and ample open space, are contributing to higher utility scores. Furthermore, the potential presence of buildings with brown-coloured bricks appears to be beneficial for the utility-based liveability score but is still a hypothesis for now. Conversely, the presence of grey hues and vehicular infrastructure, as well as vehicles themselves, negatively impacts utility scores.

To further identify the behaviour of the model's decision-making, it is necessary to analyze the neighbourhoods discussed in the beginning who scored poorly and the ones who scored better. In tables 6.3 and 6.2 the best and worst performing PC5 level zip codes were identified with their corresponding neighbourhood. For the worst performing, no clear neighbourhoods were standing out, while for the best scores, Nesselande showed a high presence in the top 10. To better understand the average behaviour of the neighbourhoods, the average utility scores per neighbourhood give vital information. Figure 6.10 shows the average utility-based liveability score per neighbourhood (zip code level 4).





**Figure 6.10:** Total average utility-based liveability score per neighbourhood (PC4 level)

Kralingse Bos, Kralingen Oost, and Nesselande are high-performing neighbourhoods, while Kralingseveer, Middelland, and Het Lage Land exhibit lower performance. Table 6.4 presents the average utility score per neighbourhood, reaffirming that Kralingse Bos, Nesselande, and Kralingen Oost are among the top-scoring neighbourhoods, whereas Kralingseveer, Het Lage Land, Struisenburg, Bospolder, and Tussendijken score lower. These locations are chosen for further analysis to analyze factors contributing to their respective performance levels.

To facilitate proper comparisons also within each neighbourhood, it is imperative to examine both well-performing and low-performing for each neighbourhood. Therefore, for Delftshaven, which has no high-performing neighbourhood, Witte Dorp will be analyzed, as it emerges as the highest-scoring neighbourhood in this vicinity. By determining these contrasting neighbourhoods, insights can be obtained into the determinants of utility-based liveability and identify areas for potential improvement.

neighbourhood	Average Utility Score	Area
Kralingse Bos	0.372902	Kralingen Crooswijk
Nesselande	0.174826	Prins Alexander
Kralingen Oost	0.039383	Kralingen Crooswijk
De Esch	-0.138599	Kralingen Crooswijk
Zevenkamp	-0.172354	Prins Alexander
Ommoord	-0.192444	Prins Alexander
Witte Dorp	-0.210963	Delftshaven
's-Gravenland	-0.225798	Kralingen Crooswijk
Spangen	-0.234818	Delftshaven
Prinsenland	-0.252413	Prins Alexander
Oud Mathenesse	-0.258310	Delftshaven
Nieuwe Westen	-0.263694	Delftshaven
Nieuw Crooswijk	-0.283179	Kralingen Crooswijk
Kralingen West	-0.303417	Kralingen Crooswijk
Oosterflank	-0.309615	Prins Alexander
Delfshaven	-0.316839	Delftshaven
Oud Crooswijk	-0.343085	Kralingen Crooswijk
Rubroek	-0.350939	Kralingen Crooswijk
Schiemond	-0.365077	Delftshaven
Middelland	-0.373266	Delftshaven
Tussendijken	-0.377501	Delftshaven
Bospolder	-0.387048	Delftshaven
Struisenburg	-0.432804	Kralingen Crooswijk
Het Lage Land	-0.458935	Prins Alexander
Kralingseveer	-0.662928	Prins Alexander

**Table 6.4:** Table with neighbourhoods, corresponding utility scores, and areas (figure 6.1 shows the location of these neighbourhoods)

In Table 6.5 the average score of each of the three areas is given.

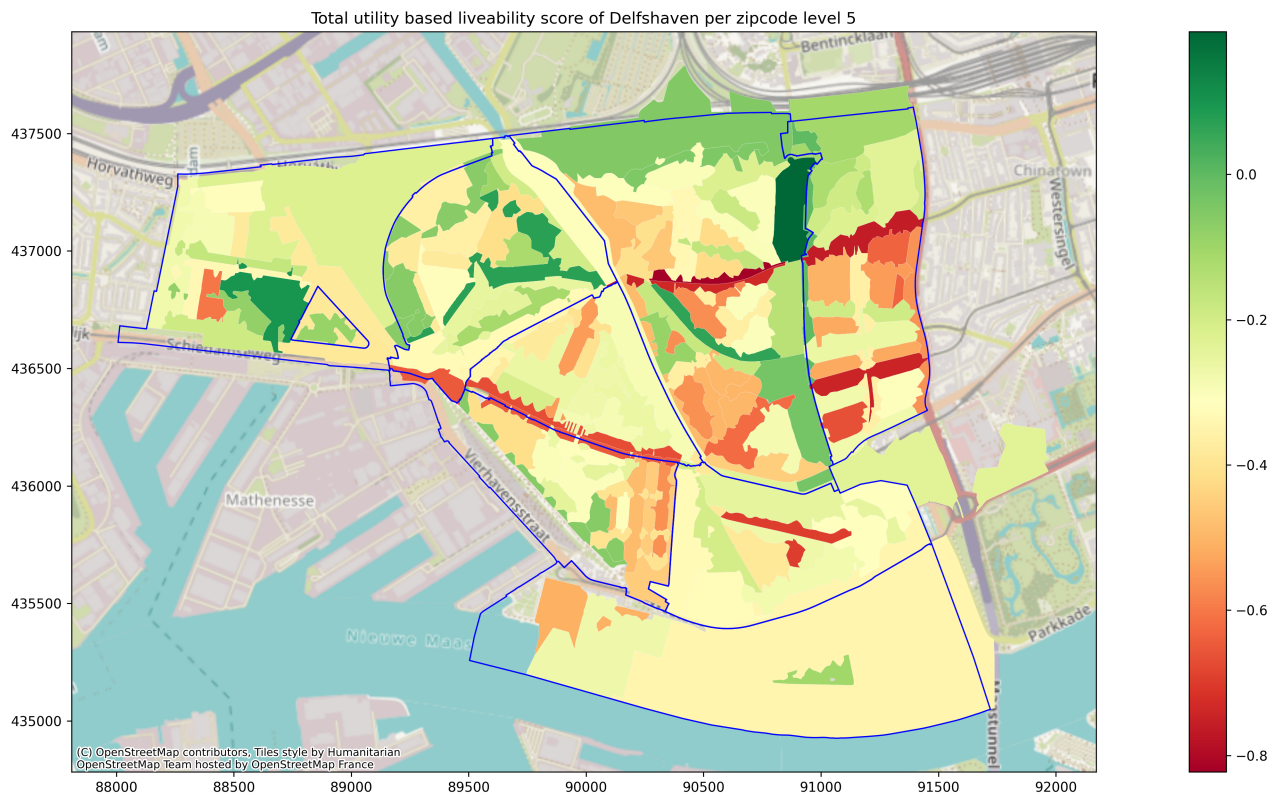
Area	Mean	Std	Min	Max
Delftshaven	-0.306955	0.501054	-2.529096	1.713997
Prins Alexander	-0.253102	0.714244	-2.808360	2.550455
Kralingen Crooswijk	-0.070684	0.657383	-2.411577	2.752395

**Table 6.5:** Summary of data on the three research areas average utility-based liveability scores

It can be seen that indeed Kralingen Crooswijk scores on average the highest and Delftshaven the lowest, while the standard deviation of Delftshaven is the lowest. This score can indicate that the neighbourhoods in Delftshaven are more similar than in the other two areas.

### 6.2.1. Delftshaven

In figure 6.11, the lowest utility scores are concentrated along long stretches of road. It's crucial to highlight that prominent among these areas are the Vierambachtstraat (indicated by the red stroke in the top right), the Nieuwe Binnenweg (red stroke below that), and the Schiedamseweg (red stroke to the left). These three roads are characterized by heavy vehicular traffic, particularly cars. As established in previous analyses, such areas are anticipated to score low in terms of liveability, given the negative impact of busy roads on the overall quality of the street view environment.



The highest-scoring average utility neighbourhood is Witte Dorp, the triangular area in the left corner, and the worst-scoring neighbourhoods are Bospolder and Tussendijken. All three will be analyzed and discussed.

### Witte Dorp

In Figure 6.12 it can be seen that Witte Dorp mainly consists of two building styles. In this style with coloured doors, it can be seen that they do not score positively unless the presence of free space is available. Otherwise, the other buildings all lay around green areas with water, scoring high. There, Witte Dorp scores high probably because of the presence of much water and trees for residential houses. In Appendix G.1 and Appendix G.2 some more random images from Witte drop can be found. Here, the same pattern occurs, but in the second photo grid it shows almost identical pictures of a tree near the waterside and in the back housing, one scored much negative, and the other positive. Also, the presence of cars and lampposts seems to be negatively influencing the score. Here, it is also worth noticing that the presence of a sidewalk seems to be negatively influencing the score as well, with the note that these sidewalks are together with parking space.



**Figure 6.12:** Random street view images with their respective utility score from Witte Dorp

### Bospolder

In Figure 6.13, the characteristics of the Bospolder neighbourhood become apparent, with a predominant emphasis on vehicular infrastructure, including roads and cars. Despite the presence of trees, indicative of some greenery, the utility score appears to be primarily influenced by the prevalence of cars and highways within the area. Appendices G.3 and G.4 offer additional random images from Bospolder, again showing the observation that in Bospolder, the abundance of roads emerges as the primary factor contributing to the low scores.



**Figure 6.13:** Random street view images with their respective utility score from Bospolder

### Tussendijken

In Figure 6.14, the characteristics of the Tussendijken neighbourhood are noticeable, showcasing a layout for vehicular transportation. Residential blocks are positioned near roads, and the presence of numerous cars further accentuates this trend. This vehicular-centric design likely contributes to the neighbourhood's negative utility score. Notably, street views without cars tend to score relatively higher, particularly when parking spaces are not visibly present. Appendices G.5 and G.6 offer additional random images from Tussendijken, reaffirming the influence of vehicular infrastructure over greenery and water features on the utility scores. In Appendix G.5 noteworthy results are observed; for instance, the 10th picture, characterized by minimal greenery and the presence of cars, scores high, as does the 12th picture, which features cars, parking spaces, and limited greenery. These observations underscore the prominent influence of vehicular elements on the neighbourhood's utility-based liveability score.



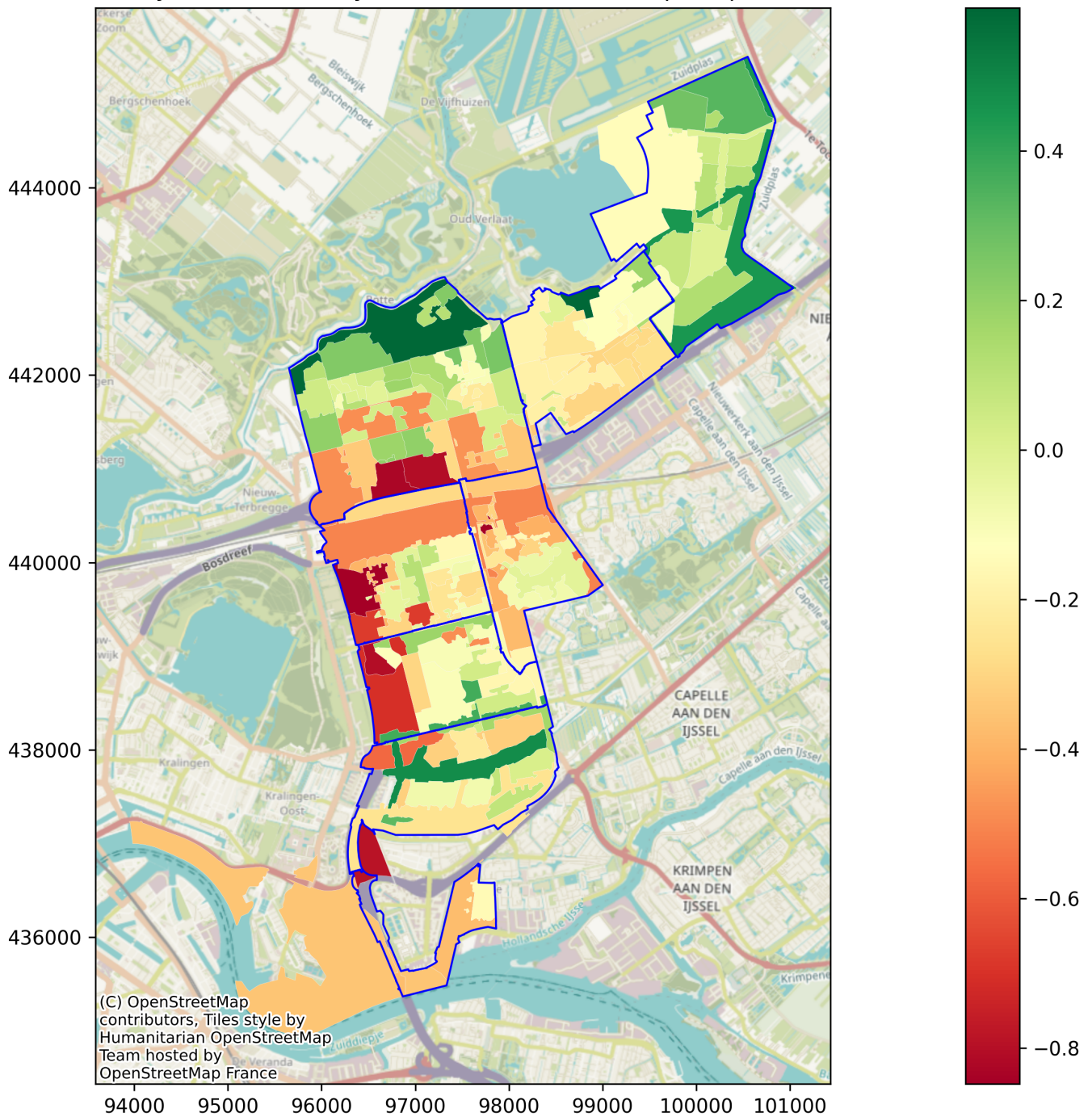
**Figure 6.14:** Random street view images with their respective utility score from Tussendijken

The analysis of Delftshaven's neighbourhoods reveals distinct patterns influencing the utility-based liveability scores and helps us understand the decision-making of the model further. From Witte Dorp (Figure 6.12), it can be found that ample free space, buildings surrounded by green areas with water score high, and street views with the presence of cars, lampposts, and sidewalks with parking spaces negatively impact the score. From Bospolder, the abundance of vehicular infrastructure emerges as an impactful negative influence on the utility score. Tussendijken also allows for this conclusion, where not only the infrastructure for cars but also cars themselves show to be negatively impacting the utility. But following this, street views without cars tend to score higher, particularly when also parking spaces are absent.

### 6.2.2. Prins-Alexander

In Figure 6.15, the average utility scores per level 5 zip code for Prins-Alexander are displayed. Notably, the lowest utility scores are concentrated near the outskirts on the right side of Kralingse Bos. Prins Alexander illustrates distinct variations in utility scores among different neighbourhoods, with some areas exhibiting no or limited zip code areas with scores tending towards the negative utility scores. Upon closer examination of the underlying geographical map, it becomes evident that neighbourhoods performing well in Prins Alexander are situated near green spaces with water nearby, consistent with previous observations indicating high utility scores associated with such features.

## Total utility based liveability score of Prins Alexander per zipcode level 5



**Figure 6.15:** Prins Alexander utility-based utility based liveability scores per zip code level 5

The highest-scoring average utility neighbourhood is Nesseland, the area in the right corner, and the worst-scoring neighbourhoods are Het Lage Land and Kralingseveer. All three will be analyzed and discussed.

#### Nesseland

In Figure 6.16, the abundance of greenery in Nesseland is evident compared to other analyzed neighbourhoods as in Delftshaven for example. The high utility scores attributed to Nesseland can probably

be attributed to the substantial presence of greenery and water features. Notably, the influence of water presence appears significant, particularly if the water is near residential housing. Appendices G.7 and G.8 further underscore this observation, as nearly every random street view image in Nesselande contains some form of greenery. The deliberate incorporation of greenery in public spaces, especially in more recently constructed houses, contributes positively to the overall liveability of Nesselande, despite the presence of roads and cars. Next to the greenery and water, the free space near residential housing also seems to be beneficial for the utility score.



**Figure 6.16:** Random street view images with their utility score from Nesselande

### Het Lage land

In Figure 6.17 the presence of roads and cars is clear. What also stands out, is the greyness of most images. These factors can explain the low utility score of Het Lage Land, while again it is clear that the presence of greenery, especially trees has a positive effect on the utility score. In Appendix G.9 and Appendix G.10 some more random images from het Lage Land can be found. Appendix G.9 especially shows the high presence of cars and roads, which have a negative influence on the utility score of Het Lage Land. In Appendix G.10 more residential homes can be seen, which provide higher utility scores in het Lage Land, with greenery present as well at these homes (less than in Nesselande), but space for cars to drive and park is evident in all of them as well.





**Figure 6.17:** Random street view images with their respective utility score from Het Lage Land

### Kralingseveer

In Figure 6.18, the negative utility score observed in Kralingseveer is primarily attributed to the extensive presence of highways and overall greyness, due to the high presence of highway views in most images. A notable discrepancy in scores is evident between street views featuring roads and those showcasing only residential houses. Appendices G.11 and G.12 further illustrate this trend, with additional random images from Kralingseveer highlighting the pervasive presence of highways. In Appendix G.11, an interesting observation emerges: residential houses with brown-coloured bricks tend to score relatively high, even in the absence of greenery. This high score could also be caused by the amount of free space without cars near the house. This suggests that factors such as the amount of free space on the street, distinct from the highway, may also influence scores, as already seen in Delftshaven. Similarly, in Appendix G.12, the influence of street features on scores becomes evident. Residential houses with brown colouring consistently score higher than those with trees, indicating that factors beyond greenery play a role in determining utility scores. Notably, subtle variations in street views, such as the presence of a red car, can result in score disparities. This can be seen mainly in Appendix G.12 (picture in the left corner) and Figure 6.18 (images 1 and 6). The difference here is only taking a step more to the right. This step results in a score which is higher than in the other street views from this place, where the difference is mainly in the presence of a red car.

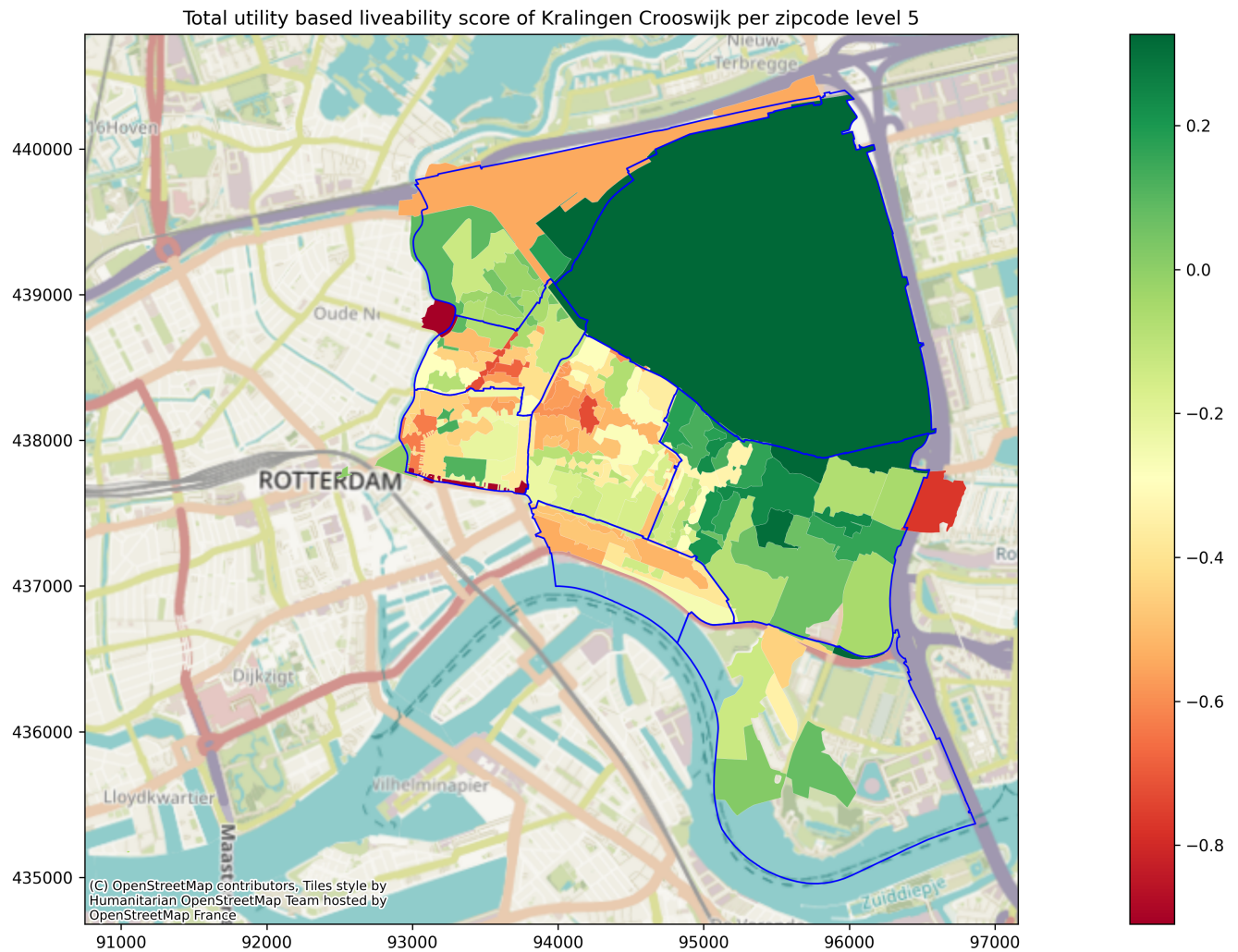


**Figure 6.18:** Random street view images with their respective utility score from Kralingseveer

From Nesselande, the positive influence of greenery and water near residential homes becomes apparent, as well as the positive influence of free space near homes. Het Lage Land shows that the presence of roads, cars, and overall greyness contributes to the low utility scores of street view images. However, the positive influence of greenery, particularly trees, becomes evident as well in the Lage Land. the presence of space for cars to drive and park appears to impact utility scores negatively. In Kralingseveer, the low utility scores are primarily attributed to the extensive presence of highways and overall greyness. However, residential houses with brown-coloured bricks tend to score relatively high, suggesting that building aspects can also influence the utility score. The positive influence of free space near houses, especially free space without cars present also becomes evident as a positive influence on the utility score in Kralingseveer.

### 6.2.3. Kralingen-Crooswijk

In Figure 6.19, the average utility scores per level 5 zip code for Kralingen-Crooswijk are displayed. Notably, the lowest utility scores are concentrated near the outskirts of Kralingen Crooswijk. As the figure shows, this area has a relatively high utility score, as much area has a positive utility score.



**Figure 6.19:** Kralingen Crooswijk utility-based liveability scores per zipcode level 5

The highest-scoring average utility neighbourhoods are Kralingse Bos and Kralingen Oost, and the worst-scoring neighbourhood is Struisenburg, which is the orange area near the water in the figure. All three will be analyzed and discussed.

### Kralingse Bos

Kralingse Bos is predominantly covered by greenery, containing a vast forest area. Therefore, the high utility score associated with Kralingse Bos is unsurprising, given the substantial influence of greenery found prior. Figure 6.20 illustrates this characteristic, showcasing abundant greenery in the neighbourhood with fewer views of residential homes. An important insight gleaned from this observation is the direct correlation between the presence of greenery in street views and higher utility scores. While this correlation may seem intuitive, it's noteworthy that when the entirety of the street view is covered in greenery, the score tends to be higher, even without any visible streets or roads. Therefore, not only the presence of greenery is beneficial, but also the area of the image covered with greenery can be influential. Appendices G.13 and G.14 further exemplify this trend, with additional random images from Kralingse Bos highlighting the positive influence of greenery. Interestingly, the influence of greenery is not limited to just trees with leaves; even trees without leaves, such as those in winter, contribute to high scores. This suggests that the model not only recognizes greenery as positive but also considers the mere presence of trees as a favourable factor in determining utility scores.



**Figure 6.20:** Random street view images with their respective utility score from Kralingse Bos

### Kralingen Oost

In Figure 6.21, analyzing the factors contributing to the high utility scores in street view images for Kralingen Oost proves challenging. While greenery and trees are present throughout the neighbourhood, they are less abundant compared to areas like Kralingse Bos. Notably, most buildings are situated near greenery or trees, yet distinguishing a clear pattern is difficult. In the figure, street views featuring trees do not consistently score higher compared to those without. For instance, in the third row, homes surrounded by greenery and more free space tend to score lower than those without such features, despite both featuring cars. An interesting observation arises from the presence of bicycles in some street views. Street views containing bicycles tend to score higher, suggesting a positive influence on utility scores. Appendices G.15 and G.16 further highlight this trend, with additional images from Kralingen Oost showing similar effects of bicycles on scores. In the top right corner of Appendix G.15, the presence of bikes seems to have a positive influence on the utility score, while in the top left street view, there also appear to be bicycles, but the score does not tend to be positive, perhaps this is due to only the rear of bicycles being visible on the street views. Surprisingly, even a clear bicycle path, without bicycles on it, tends to have a negative influence on utility scores, particularly when greenery is present near residential homes. This indicates that the model may only capture the direct presence of bicycles rather than indirect indicators like bicycle lanes. Additionally, the presence of a bricked sidewalk in front of homes appears to positively influence utility scores, with street views featuring only the sidewalk and the home scoring higher in Kralingen Oost. This would suggest that the closeness of the car capturing the images to the homes is also affecting the utility scores, as capturing the road in the images tends to lead to lower scores in comparison to only capturing the home or the home and the sidewalk.



**Figure 6.21:** Random street view images with their respective utility score from Kralingen Oost

### Struisenburg

In Figure 6.22, the dense urban layout and towering buildings of Struisenburg are apparent, leaving little space for greenery. Many areas in Struisenburg are dedicated to parking and roads, possibly resulting in numerous street views with low utility scores. Once again, the presence of water and trees emerges as beneficial for utility scores, while the presence of cars decreases the utility score. Appendices G.17 and G.18 further illustrate these trends, with additional images from Struisenburg reinforcing the positive influence of water and greenery on utility scores. In particular, the combination of water and greenery significantly boosts utility scores, as seen in the first two images of the third row in Appendix G.17. Conversely, the presence of cars tends to decrease utility scores, potentially by limiting free space in street views, as observed in the last image of Appendix G.18.



**Figure 6.22:** Random street view images with their respective utility score from Struisenburg

From Kralingse Bos, the positive influence of greenery becomes more apparent. It also showed that trees without green leaves are still influential and that just the presence of the tree, with or without leaves, has a positive effect on the utility score. Kralingen Oost showed the positive influence of the presence of bicycles in street views, while bicycle lanes negatively influenced the utility score. Next to this, bricked sidewalks emerged as positively influencing the utility scores. From Struisenburg it became clear that neighbourhoods built for vehicle mobility and parking space score low. Due to this building type of this area, little space is available for greenery and water. However, when present in Struisenburg, positively influences the utility scores. Also, the presence of cars tends to decrease utility scores.

### 6.3. Face validity conclusions

The face validity analysis across various neighbourhoods highlights several key factors which tend to influence the utility score. The presence of greenery, including trees with or without leaves, consistently seems to contribute to higher utility scores. Conversely, areas characterized by limited greenery due to dense urban layouts or extensive vehicular infrastructure tend to score lower. Additionally, the presence of bicycles in street views seems to positively impact utility scores, while bicycle lanes without the presence of bikes seem to have a negative influence. The presence of water seems to also be beneficial for the utility score, especially in combination with nearby greenery. The availability of free space, particularly near residential areas, emerges as a significant factor positively affecting utility scores. Conversely, the presence of cars and vehicular infrastructure, such as roads and highways, but also parking spaces and lampposts, seems to decrease utility scores. Specifically, street views without cars tend to score higher. Furthermore, the presence of specific architectural features, such as bricked

sidewalks and buildings with brown bricks, seems to result in higher utility scores. In general, the amount of colour grey in the street view also seems to decrease the utility score. Most of these findings are in accordance with what was found in the literature review about factors influencing liveability, but factors such as bricked sidewalks, brown bricks, and greyness were not found in the literature. This conclusion answers the third sub-question: 'What face validity conclusions can we draw about the decision behaviour of the CVDCM without the application of XAI?'

## 6.4. Result validation

To validate the results obtained (the utility-based liveability score), 'the Wijkprofiel' of the municipality of Rotterdam can be used. Here, the neighbourhoods of Rotterdam are analyzed in terms of their 'liveability', which is measured using various subjective and objective factors. This data is at the neighbourhood level, PC4. This allows for correlation analysis between their and our data. For the Wijkprofiel, 3 indicators are measured, including all subjective and objective factors (Social, Physical and Safety). The subjective measurements in the Wijkprofiel are taken via survey measurements. The objective measurements are calculated by own measurements of the Municipality. The information in this paragraph about the Wijkprofiel is all directly from Rotterdam [115].

The following factors have been used for correlation analysis: 'Social Index', 'Social Index - subjective', 'Social Index - objective', 'Judgement quality of life', 'Physical Index', 'Physical Index - subjective', 'Physical Index - objective', 'Housing experience', 'Housing subjective', 'Public Space subjective', 'Amenities subjective', 'Environment subjective', 'Housing objective', 'Public Space objective', 'Amenities objective', 'Environment objective', 'Safety Index', 'Safety Index - subjective', 'Safety Index - objective', 'Safety Experience'.

The subjective score consists of indicators that come from survey questions, covering opinions (such as confidence in the government) and ratings (such as satisfaction with housing). The objective score includes indicators derived from various registrations or factual data from the survey questions, asking citizens about facts (such as educational attainment) or behaviours (such as participation in social activities).

- Social index  
The social index contains information on themes such as perception of quality of life, self-reliance, co-efficacy, participation, and bonding, providing an average of the subjective and objective measurements.
- Social index - objective  
This is equal to the Social index but includes only the average of the objective measurements.
- Social index - subjective  
This is equal to the Social index but includes only the average of the subjective measurements.
- Judgement quality of life  
This is a picture of people's assessment of their quality of life. It serves as a summary perception score for the subjective dimension of the Social Index. It reflects how satisfied residents are with the quality of their lives, considering factors such as activities outside the home, contacts with family and friends, health, and well-being.
- Physical index  
The physical index contains information on themes such as the living experience, housing, public space, facilities, and environment, providing an average of the subjective and objective measurements.
- Physical index - objective  
This is equal to the Physical index but includes only the average of the objective measurements.
- Physical index - subjective  
This is equal to the Physical index but includes only the average of the subjective measurements.
- Housing experience  
This gives a general picture of living in a neighbourhood. Satisfaction with the housing situation indicates how pleasant people find living in the neighbourhood and largely determines whether they

plan to move soon. Neighbourhoods, where people live satisfactorily and do not want to leave, have a stronger market position than those where this is less the case. It includes residential satisfaction and propensity to move.

- **Housing subjective**  
This shows the residents' satisfaction with their homes and the attractiveness of the built environment in their neighbourhood. It includes satisfaction with various housing aspects, such as size, layout, type, size of storage and outdoor space, indoor climate (insulation and ventilation), safety of the dwelling, price-quality ratio, and the attractiveness of the buildings in the neighbourhood.
- **Housing objective**  
This concerns the quality of housing and building stock in the neighbourhood. It includes housing stock quality (such as the share of 'vulnerable' dwellings, WOZ value, and risk of foundation problems), occupancy of housing stock (such as vacancy and overcrowding), popularity of housing supply (interest in rental properties and speed of sale of owner-occupied properties), and state of maintenance of own home, neighbouring properties, and buildings in the neighbourhood as assessed by residents.
- **Public space objective**  
This portrays the quality of public space and traffic safety in the neighbourhood. Since 2019, the national CROW measurement method has been used, to determine the quality level in the neighbourhood through random inspections. The theme includes cleanliness (litter, waste bins, excrement, defacement, and weeds), integrity (paving and public greenery), and traffic safety (traffic accidents).
- **Public space subjective**  
This represents residents' ratings of the quality of public space and road safety. It includes assessments of cleanliness (rubbish on the street, dirt next to the container, destruction of street furniture and bus/tram shelters), greenery and water (availability and attractiveness), pavement maintenance (pavements and cycle paths), traffic safety (safety of pavements and cycle paths, traffic behaviour, and collisions), satisfaction with neighbourhood accessibility to car traffic, and satisfaction with street lighting quality.
- **Amenities objective**  
It provides a picture of the proximity of various facilities in the neighbourhood. It considers the distance to facilities and the number of facilities within reach. Neighbourhoods with many facilities nearby score higher. This includes shops for daily shopping, sports and play facilities, educational facilities, primary care facilities, public transport, and issues like vacancies in shops.
- **Amenities subjective**  
This theme examines residents' satisfaction with the provision of facilities and their assessment of the availability of various types of facilities in and around their neighbourhoods. It includes the overall assessment of facilities, the presence of daily facilities, primary care facilities, sports facilities, educational facilities, public transport, and parking facilities.
- **Environment objective**  
This relates to the quality of the environment in the district, using data from the DCMR Environmental Department. It includes air quality (NO<sub>2</sub> concentration) and noise quality (average noise level).
- **Environment subjective**  
This theme provides a picture of the quality of the environment as experienced by residents, including the extent to which they experience odour, noise, and water nuisance. It includes odour nuisance from traffic, activity, sewerage, and water, noise nuisance from traffic, industry, and construction/demolition activities, and water nuisance in gardens or courtyards and under dwellings.
- **Safety index**  
The safety index contains information on themes such as safety perception, theft, violence, burglary, vandalism, and nuisance, providing an average of the subjective and objective measurements.
- **Safety index - subjective**  
This is equal to the Safety index but includes only the average of the subjective measurements.



- **Safety index - objective**  
This is equal to the Safety index but includes only the average of the objective measurements.
- **Safety experience**  
This theme provides an overall picture of residents' perceived safety in the neighbourhood. It includes satisfaction with living in the neighbourhood, avoidance behaviour (such as not opening the door at night or avoiding places in the neighbourhood rated as 'unsafe'), and the extent to which residents think they or someone else in their household is at risk of becoming a victim of burglary, pickpocketing, street robbery, or assault.

These factors in one way or another could be argued that they are influential on the utility-based liveability score. Especially factors such as the physical index, the housing experience, the housing factor, and the public space factor. This makes it interesting to investigate the Pearson correlation of these factors with the utility-based liveability score. The Pearson correlation is a statistical measure that evaluates the strength and direction of the linear relationship between two continuous variables (the utility-based liveability score and a Wijkprofiel factor). It ranges from -1 to 1, where +1 indicates a perfect positive linear relationship, meaning as one variable increases, the other variable also increases in a perfectly linear manner. A value of 0 indicates no linear relationship, meaning changes in one variable do not predict changes in the other variable. A value of -1 indicates a perfect negative linear relationship, meaning as one variable increases, the other variable decreases in a perfectly linear manner (Pearson [107]). The p-value associated with each correlation measures the probability that the observed correlation could have occurred by random chance. A p-value below 0.05 typically indicates statistical significance, meaning the observed correlation is unlikely to be due to random chance.

To analyse the correlation between the two data sources, the data needs to be identically distributed to provide useful results. The Wijkprofiel data ranges from the interval of 0 until 200, and the utility-based liveability score from -4 to 4, therefore, both datasets need to be set to equal distribution and interval. Therefore, the StandardScaler from the 'scikit-learn' Python package is used which sets both data frames to be Normally distributed with a mean of 0 and a standard deviation of 1. Table 6.6 shows the correlation coefficient of each factor with the utility-based liveability score.

It's important to note that our correlation calculations are based on a relatively small dataset, encompassing approximately 25 neighbourhoods. With only 25 data points for each correlation calculation, the potential influence of outliers and biases is heightened. This limited sample size also impacts the statistical significance of our findings, emphasizing the need for further validation with larger and more diverse datasets to strengthen our conclusions.

Wijkprofiel Data	Correlation with average utility-based liveability score on neighbourhood PC4 Level	p-value
Amenities - subjective	0.550567	0.005
Social Index - subjective	0.432637	0.035
Physical Index	0.417881	0.042
Environment - subjective	0.387550	0.061
Physical Index - subjective	0.386491	0.062
Social Index	0.379906	0.067
Housing - objective	0.376814	0.070
Judgement Quality of Life	0.365884	0.079
Housing - subjective	0.347488	0.096
Public Space - subjective	0.277266	0.190
Housing Experience	0.262041	0.216
Safety Experience	0.260522	0.219
Social Index - objective	0.211070	0.322
Physical Index - objective	0.133619	0.533
Environment - objective	0.123754	0.565
Safety Index - subjective	0.049564	0.818
Safety Index	-0.005382	0.980
Amenities - objective	-0.051541	0.811
Safety Index - objective	-0.107128	0.618
Public Space - objective	-0.210252	0.324

**Table 6.6:** Correlation of Wijkprofiel Data with Average Utility Score of neighbourhoods

The analysis of the Pearson correlations between the utility-based liveability score and various factors from the Wijkprofiel dataset (Table 6.6) reveals several interesting patterns and insights. The highest observed correlation is between the utility-based liveability score and subjective perceptions of amenities, with a correlation coefficient of 0.550567 and a p-value of 0.005. This significant positive correlation suggests that as subjective evaluations of neighbourhood amenities improve, so does the utility-based liveability score.

Other subjective factors also show notable correlations, although weaker than amenities. The social index (subjective) has a correlation of 0.432637 with a p-value of 0.035, and the physical index (subjective) shows a correlation of 0.386491 with a p-value of 0.062. These findings indicate that subjective perceptions of social and physical attributes of neighbourhoods are positively associated with utility-based liveability scores. The moderate correlation of 0.417881 (p-value 0.042) with the physical index suggests it also plays a role in influencing the liveability score.

Interestingly, objective factors generally exhibit lower correlations with the utility-based liveability score. For instance, the physical index (objective) and social index (objective) show correlations of 0.133619 (p-value 0.533) and 0.211070 (p-value 0.322), respectively. These weak correlations, coupled with high p-values, suggest that objective measurements of physical and social conditions in neighbourhoods have less influence on the liveability score compared to subjective evaluations.

The weak correlations observed for many objective factors, such as the safety index (objective) with a correlation of -0.107128 (p-value 0.618), indicate that these objective measurements do not align closely with the utility-based liveability score. This discrepancy highlights the potential differences between measured conditions and perceived experiences, emphasizing the importance of considering subjective assessments in evaluating neighbourhood liveability.

The correlation results indicate a general trend where subjective evaluations tend to correlate more strongly with the utility-based liveability score than objective measurements. This pattern underscores the significance of residents' perceptions and experiences in assessing neighbourhood quality and liveability.

Given the relatively small dataset of approximately 25 neighbourhoods, the potential influence of outliers and biases is heightened, and the statistical significance of the findings is limited, as most p-values are above 0.05. Therefore, further validation with larger and more diverse datasets is necessary to strengthen these conclusions.

Regarding the face validity analysis identifying greenery and water as possibly highly influential, it is interesting to analyze the correlations for these factors as well. The Wijkprofiel data does not provide direct features which measure greenery or water presence, therefore it is worthwhile to explore the correlation between green amenities data and the utility-based liveability score. The municipality's green data at the neighbourhood PC4 level includes the following data:

- Total Surface Area: Overall neighbourhood area, including all land and water.
- Water Area: Total area covered by water.
- Land Area: Total land area, including green and built-up areas.
- Green Space Area: Area covered by trees, shrubs, woods, and grass.
- Percentage of Green Space: Proportion of the total area that is green.
- Percentage of Green Space Relative to Land: Proportion of land area that is green.
- Percentage of Vegetated Terrain: Percentage of terrain areas that are green.

The correlation results are given in table 6.7

Green data	Correlation with average utility score on neighbourhood PC4 level	p-value
Area surface	0.692084	0.0001
Water area	0.688032	0.0001
Green area	0.677486	0.0002
Land area	0.604979	0.0014
Percentage vegetated terrain area	0.568442	0.0030
Percentage of amount green on land area	0.553046	0.0041
Percentage of green on total area	0.490012	0.0129

**Table 6.7:** Correlation of greenery data with average utility score of neighbourhoods

The analysis of the correlations between green and water data and the utility-based liveability score provides valuable insights into the impact of environmental factors on neighbourhood liveability. Table 6.7 presents the correlation coefficients and p-values for various green and water-related features.

The strongest correlation is observed between the total surface area and the utility-based liveability score, with a correlation coefficient of 0.692084 and a highly significant p-value of 0.0001. This suggests that larger neighbourhoods tend to have higher utility-based liveability scores. The large area may provide more space for green amenities, recreational areas, and water features, contributing positively to the perceived liveability.

The water area also shows a strong positive correlation of 0.688032 with a p-value of 0.0001. This indicates that neighbourhoods with larger water bodies tend to have higher liveability scores. The total green area has a significant positive correlation of 0.677486 with a p-value of 0.0002. This underscores the importance of green spaces in enhancing neighbourhood liveability. The correlation between land area and the utility-based liveability score is also substantial, with a coefficient of 0.604979 and a p-value of 0.0014. This suggests that larger land areas, which can accommodate more green spaces and amenities, are associated with higher utility-based liveability scores. However, it is slightly lower than the correlations for total area, green area, and water area, indicating that the presence of green and water features might be more influential than just the land area. The percentage of vegetated terrain area shows a positive correlation of 0.568442 with a p-value of 0.0030. This indicates that a higher proportion of vegetated areas within the terrain contributes to better utility-based liveability scores.

The analysis demonstrates significant and logical correlations between the utility-based liveability score and various green and water-related factors. The strong positive correlations with total area, water area, and green area reinforce the intuitive understanding that both greenery and water positively influence neighbourhood liveability. Larger areas with green and water features are perceived as more liveable, likely due to the aesthetic, recreational, and health benefits they provide, something we have seen in the analysis of Kralingse Bos for example.

#### 6.4.1. Colour influence

Face validity results suggest that the colour green is commonly linked with higher utility scores, likely due to the association with natural greenery. Conversely, grey tends to be associated with lower utility scores. Therefore, the analysis focuses on the influence of colours alone. PNG images composed solely of these colours are inputted into the model to calculate the utility-based liveability score specific to each colour. These images do not depict street views but are uniformly coloured images. Table 6.8 displays the utility scores corresponding to images consisting solely of individual colours.

Colour	Utility score
Black	-0.61
Blue	-1.07
Brown	-0.34
Green	-1.28
Grey	-0.13
Orange	-0.63
Pink	-0.38
White	-0.15
Yellow	-0.88

**Table 6.8:** Utility score of different colours, where each colour is represented by a PNG image of only that colour

The analysis of utility-based liveability scores for uniformly coloured PNG images reveals intriguing insights. Contrary to expectations, the colours green and blue, which are typically associated with natural elements like greenery and water, exhibit the lowest utility scores, -1.28 and -1.07 respectively. This challenges the conventional belief that these colours inherently contribute to higher liveability scores. The lower score for green suggests that the actual semantic content of greenery, such as the presence of trees, parks, and natural landscapes, plays a critical role in enhancing utility-based liveability, rather than the colour itself. Similarly, the low score for blue indicates that water's contribution to liveability is not adequately captured by the colour alone. Water bodies in real urban environments present a complex visual and experiential context that a single blue colour cannot represent.

These findings emphasize the importance of considering not just the colour itself but also the broader context and meaning that these colours represent. While green and blue are associated with natural elements, their impact on liveability is significantly influenced by how these elements are integrated and perceived within the urban landscape. This underscores the necessity for a nuanced approach when evaluating environmental factors in urban liveability assessments.

## 6.5. Validation conclusion

The face validity study and data validation collectively enhance our understanding of the factors possibly influencing the utility-based liveability score. The strong correlations between greenery, water presence, and utility scores underscore the significant role these natural elements play in urban liveability. Specifically, the positive correlations with green areas and water bodies confirm that these features contribute to higher liveability scores, aligning with findings from the perceived liveability literature review that emphasize the benefits of green and blue spaces in urban environments.

However, the analysis of uniformly coloured images reveals that the semantic context of natural elements is crucial. The unexpectedly low scores for green and blue colours suggest that merely the

presence of these colours does not suffice, the experience and perception of greenery and water are essential for enhancing liveability. This finding is again supported by literature that highlights the multi-faceted impact of green and blue space.

Furthermore, the data validation against the neighbourhood profile indicates that many objective factors do not show strong correlations with the utility-based liveability score. This discrepancy highlights the potential divergence between subjective perceptions and objective measurements of neighbourhood conditions. The subjective evaluations of social and physical indices tend to correlate more strongly with utility-based liveability scores, emphasizing the importance of residents' perceptions and experiences in assessing urban quality. This aligns with literature emphasizing the significance of subjective well-being and residents' satisfaction in evaluating urban liveability.

In conclusion, the validation studies underscore the nuanced interplay between environmental elements, subjective perceptions, and objective measurements in determining urban liveability. The strong correlation between greenery and water with utility-based liveability scores reaffirms their importance, while the findings on colour highlight the necessity of considering the broader semantic context.

# 7

## Application of LIME: case studies

The application of LIME in this chapter builds upon the methodology and implementation chapter to conduct detailed case studies on the performance of LIME. In the actor analysis chapter, the expert panel is discussed. This panel will in practice be the actors to discuss the quality and interpretation of these explanations. The previous chapter's face validity analysis predominantly found that greenery, water, and free space positively influence the liveability score, whereas vehicle infrastructure negatively impacts it. This chapter investigates whether LIME, not only confirms these conclusions but also provides additional insights. By comparing the decision behaviour interpretations with and without LIME, the analysis aims to determine the added value of explainable AI in enhancing the interpretability of CVDCMs and supporting informed decision-making in the municipality of Rotterdam. Given LIME's ability to highlight important features (superpixels) within an image, it is particularly valuable for policymakers to identify elements contributing to low utility scores. Analyzing these features can provide insights for potential policy recommendations. These possible elements are guided by the findings on influential factors on urban liveability from the literature review. The complete methodology is applied to these street-view images, with detailed results presented. Multiple case studies are discussed to strengthen the findings. In total, 7 cases are shown. While this may seem a small sample size, it is important to note that these cases were selected because they represent a diverse set of street-view images and utility score outcomes. Limiting the case studies to 7 was based on the observation that most results exhibit similar patterns and behaviours (the same results were seen in other cases as well). Therefore, including more cases would likely not provide additional insights and could overwhelm the reader with redundant information. It is also acknowledged that more case studies are expected for a comprehensive analysis. However, given the consistency of the results observed in these selected cases, the chosen number of case studies is sufficient to demonstrate the methodology's application and understanding of the model's outputs.

The first case study is presented in detail, illustrating every analysis step. Only the key results are highlighted for the remaining case studies, focusing on the important findings. This approach ensures clarity and conciseness while effectively conveying the necessary information.

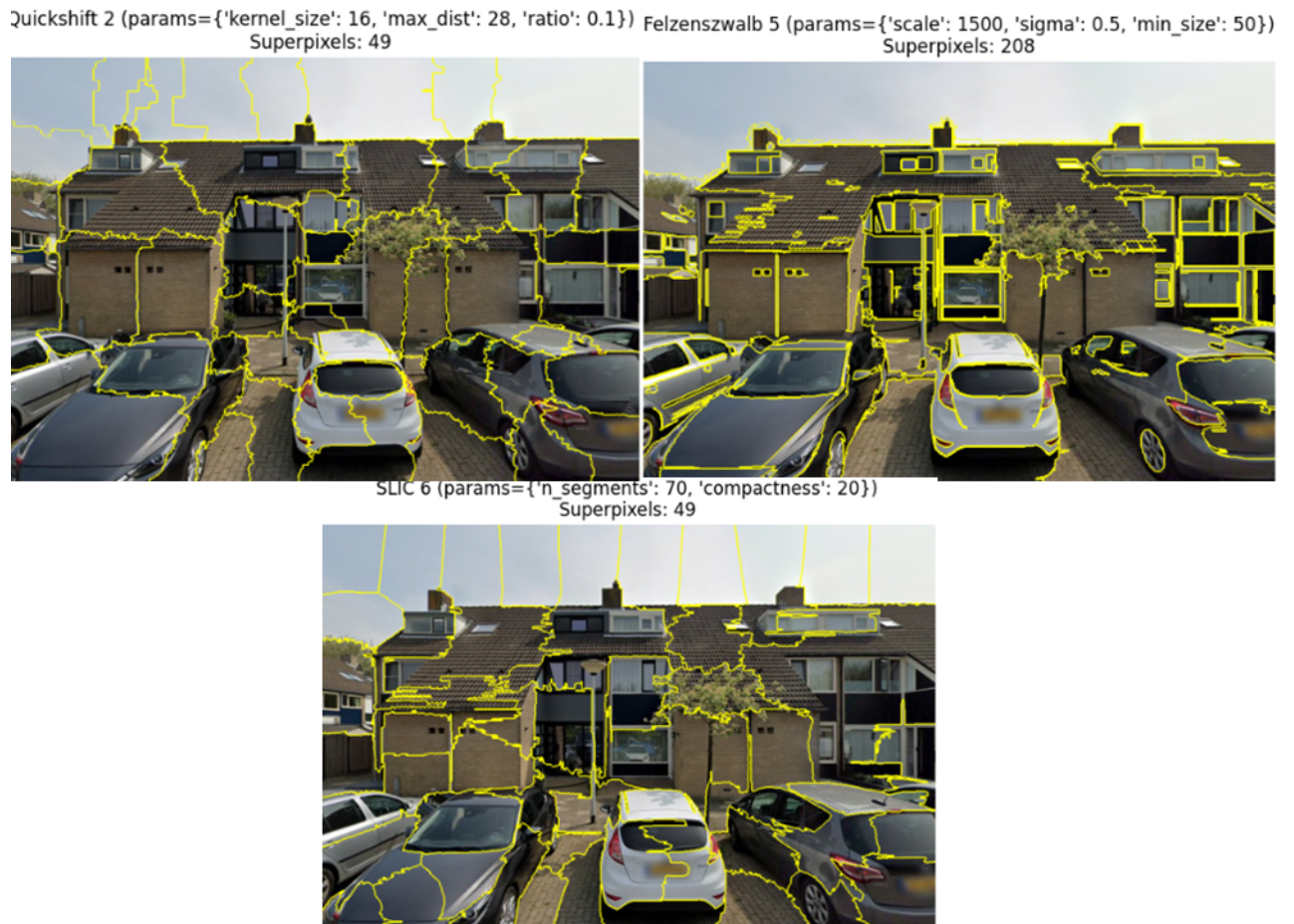
### 7.1. Case study 1

The image in figure 7.1 is selected to demonstrate the LIME image pipeline described in the methodology and implementation section.



**Figure 7.1:** Street view image used in case study 1, with a utility score of -1.06

The first step is to create segmentations of this image. For each of the three segmentation algorithms (Quickshift, Felzenszwalb, and SLIC), eight possible segmentations are generated. The visualizations of these segmentations can be found in Appendix I, where the number of superpixels and the parameters used for each segmentation are also documented. Based on visual inspection, the modeller suggests the following segmentations as the most well-suited for each of the three algorithms.



**Figure 7.2:** Three different segmentation algorithms used for the street view image

The three segmentation algorithms, Quickshift, Felzenszwalb, and SLIC, each provide some level of semantic meaning, though they do so in different ways. Felzenszwalb tends to produce a higher number of superpixels on average, capturing more detailed information in their segmentations. A ground truth image is used to scientifically determine the best segmentations. Figure 7.3 shows the user-created ground truth image of the street view. This ground truth image serves as a reference for evaluating the accuracy and effectiveness of the segmentations generated by each algorithm.





**Figure 7.3:** Modeller created ground truth image of the street view image

Here, clear objects such as cars, pavement, sidewalks, homes, lampposts, greenery, and sky are visible. According to the modeller, the segmentation algorithms should effectively segment these objects to provide semantically meaningful segmentations. Table 7.1 presents the ground truth metric results for the three selected segmentation algorithms. For a comprehensive view, Appendix I includes the entire table for each segmentation algorithm. Each value in the table is rounded to two decimal places to ensure clarity and precision.

Segmentation	ARI	Jaccard	Dice	Vol
Quickshift 2	0.01	0.75	0.86	[5.08 0.50]
Felzenszwalb 5	0.03	0.60	0.76	[5.07 0.34]
SLIC 6	0.01	0.79	0.88	[5.15 0.49]

**Table 7.1:** Metric results of the analysis of the ground truth image with the 3 used segmentation algorithms

The table shows that in terms of Jaccard and Dice metrics, Felzenszwalb scores lower than SLIC and Quickshift. For the other two metrics, Adjusted Rand Index (ARI) and Variation of Information (Vol), the scores are approximately equal across all three segmentation algorithms. The high Dice and Jaccard values for Quickshift and SLIC indicate that these algorithms align well with the modeller-defined ground truth segmentation.

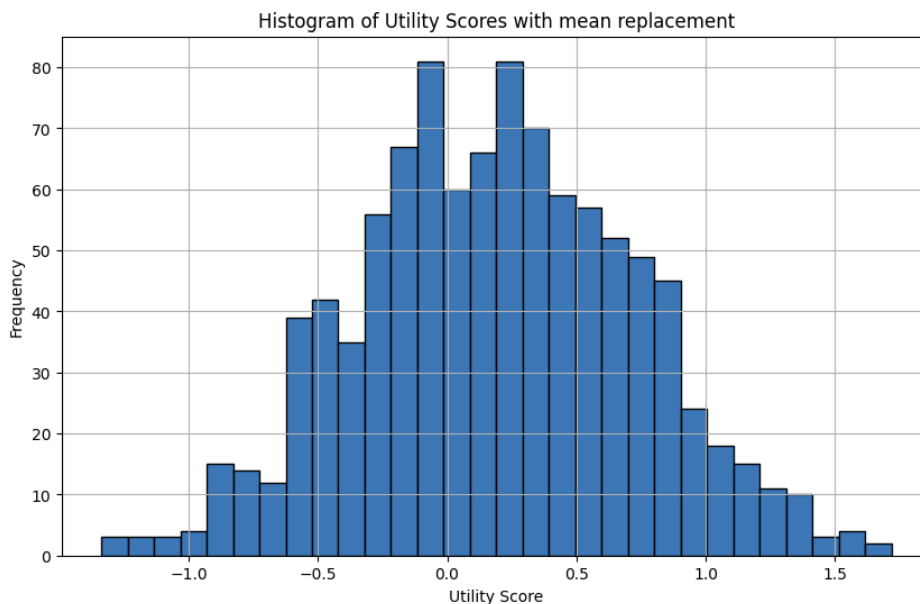
The ARI metric reflects the similarity between the clusters (segments) created by the algorithm and the ground truth, with a score close to zero indicating that the segments are not significantly better than random assignments. The Jaccard index, which measures the intersection over the union of the segments with the ground truth, is notably higher for Quickshift and SLIC, demonstrating their effectiveness in correctly identifying similar regions. The Dice coefficient shows similar results with high values for Quickshift and SLIC, showing their quality in segmenting the images in a manner consistent with the ground truth.

The Vol metric consists of two parts: the split and merge components, indicating the amount of over-segmentation and under-segmentation, respectively. The scores suggest that all three algorithms have comparable performance in terms of over-segmenting and under-segmenting the image, with Quickshift showing a slightly better balance.

The significantly lower ARI compared to the Jaccard and Dice scores may indicate that the segmentation algorithms and the ground truth do not perfectly match, even though the actual segmentation boundaries are similar. This discrepancy can occur because ARI is sensitive to the exact segments, while Jaccard and Dice focus more on the overlap between segments, regardless of specific labels. While this lower ARI might suggest that the segmentation algorithms are not perfectly aligning their labels with the ground truth, it does not necessarily diminish their use for generating meaningful superpixels. The high Jaccard and Dice scores are more indicative of the algorithms' ability to correctly identify and segment similar regions in the image.

Given these results and considering that Quickshift is the standard segmentation method used in LIME, Quickshift is chosen for creating superpixels for this image. Despite the slight differences in metric scores between Quickshift and SLIC, Quickshift is deemed to create the most semantically meaningful segmentation. This choice is driven by the higher Jaccard and Dice scores, indicating a closer match to the ground truth segmentation.

The analysis of the distribution of perturbed images determines the deterministic classification threshold for the LIME explanation. Figure 7.4 displays the distribution of utility scores obtained from sampled perturbed images using a Bernoulli(0.5) distribution for this particular image.

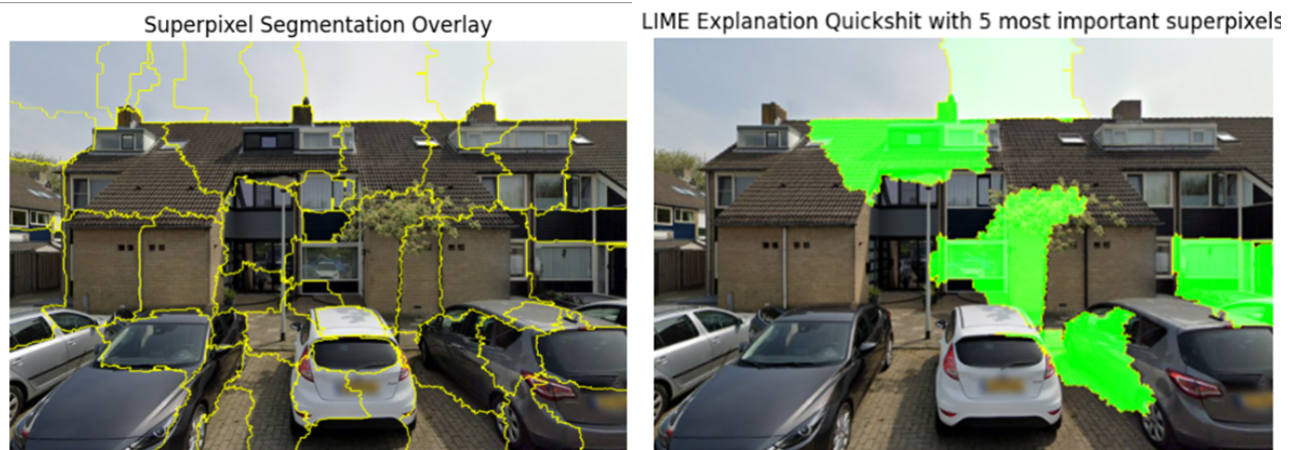


**Figure 7.4:** Distribution of the utility scores of the sampled perturbed images from the original street view image

In this perturbation sampling run, the mean utility score is 0.19, with a standard deviation of approximately 0.5. This mean score serves as the classification threshold of 0.19, which is used in creating the LIME explanation. Consequently, the classification 'bad' is defined by the interval  $(-\infty, 0.19)$ . To facilitate the LIME explanation, each image with a utility score falling within this interval is considered to belong to the 'bad' class, while those outside this range are classified as 'good'. Given that the original image has a utility score of -1.06, it falls within the 'bad' classification category. Thus, the LIME explanation focuses on explaining why this image is classified as 'bad' and identifies which regions of the image contribute positively and negatively to this classification. Superpixels with positive weights signify their contribution to the 'bad' classification, whereas those with negative weights contribute to the 'good' classification. It is important to note that the utility score of the original image no longer determines the importance in this context; rather, it serves solely to define the class ('bad' or 'good') that the LIME explanation aims to interpret and explain. The explanation therefore is not about what features are beneficial to the score of the original image, but to the classification of the whole interval the classification of the original image is from.

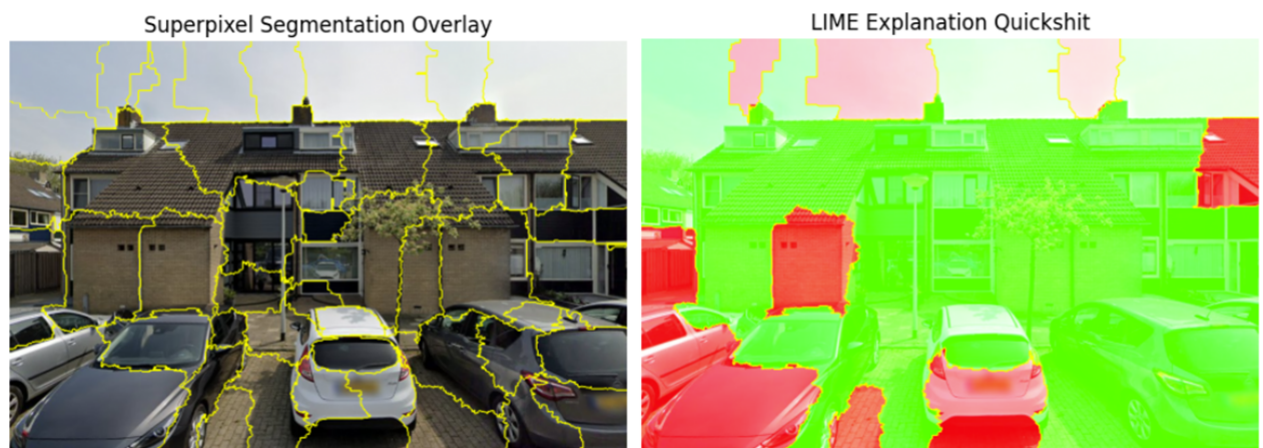
In the following analysis, first, the deterministic classification function is used. After that, the probabilistic classification function is used.

The LIME explainer function was made with 1000 samples using the Quickshift segmentation algorithm. From these samples, 490 were classified as 'bad' and 510 as 'good', resulting in a Binary Classification Ratio of 0.49 (490 'bad' samples divided by 1000 total samples), indicating sufficient data for training the LIME linear model. Figure 7.5 presents the results for the 5 most important superpixels based on their absolute weights.



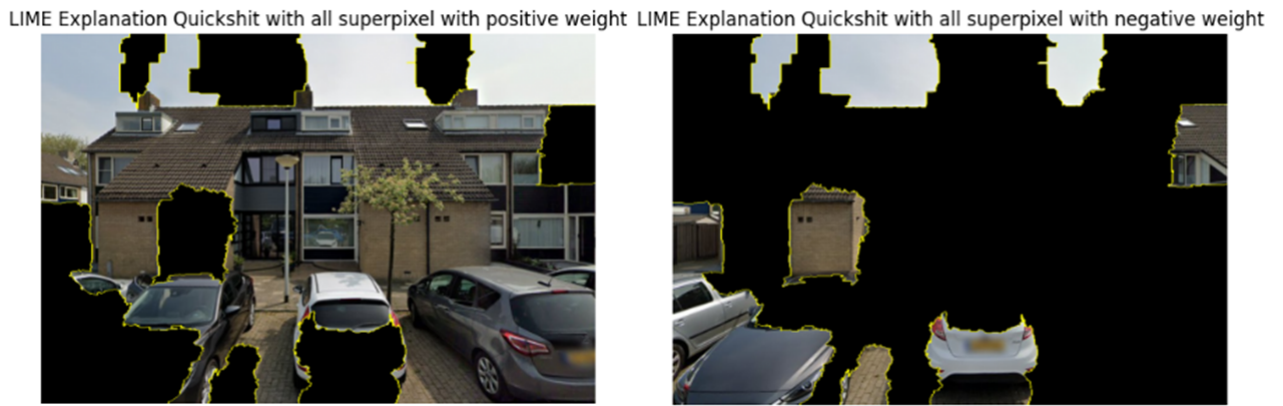
**Figure 7.5:** The 5 most important (highest absolute superpixel weight) superpixels of the LIME explanation with threshold 0.19 for the deterministic classification.

From the figure, it is evident that the 5 superpixels with the highest absolute weights all positively influence the 'bad' classification. However, these superpixels are part of a larger semantic object, making it challenging to pinpoint their semantic meaning in isolation. To gain a broader perspective, Figure 7.6 displays the influence of all superpixels in the image.



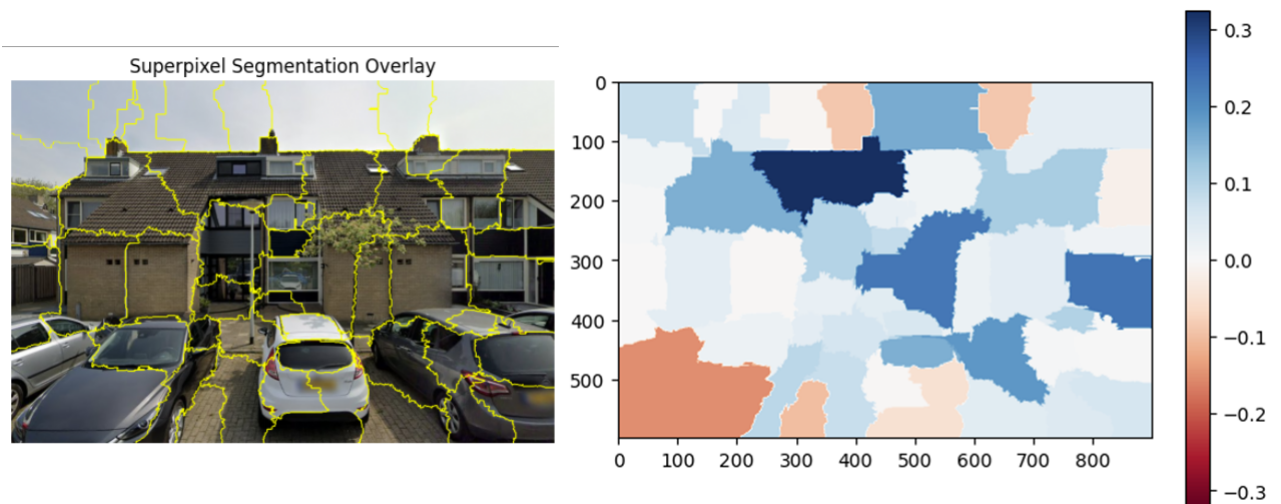
**Figure 7.6:** Influence (positive or negative) of all the superpixels of the LIME explanation with threshold 0.19 for the deterministic classification

Similarly, interpreting semantic meaning here is complex, for example, parts of cars show both positive and negative weights. Overall, a majority of the image contributes to the 'bad' classification, indicated by the high presence of red (negative) and lesser green (positive) superpixels. To provide a clearer distinction, Figure 7.7 separates superpixels into positive-weighted (left) and negative-weighted (right) groups.



**Figure 7.7:** On the left, only superpixels with positive weight are visible, and on the right, only superpixels with negative weight are visible.

This division shows that the LIME explanation does not straightforwardly associate specific objects in the street view with contributing to the 'bad' classification. For instance, parts of buildings and cars influence both positively and negatively the classification. Notably, the lamppost and small tree are the only identifiable objects among the positively weighted superpixels. However, since these objects are not accurately segmented, their contribution to positive superpixel weights is compromised. The weights of the superpixels vary widely, with higher absolute values indicating greater importance to the classification. Figure 7.8 presents a heatmap illustrating the distribution of these weights.



**Figure 7.8:** Heatmap of the superpixels weights for the deterministic classification

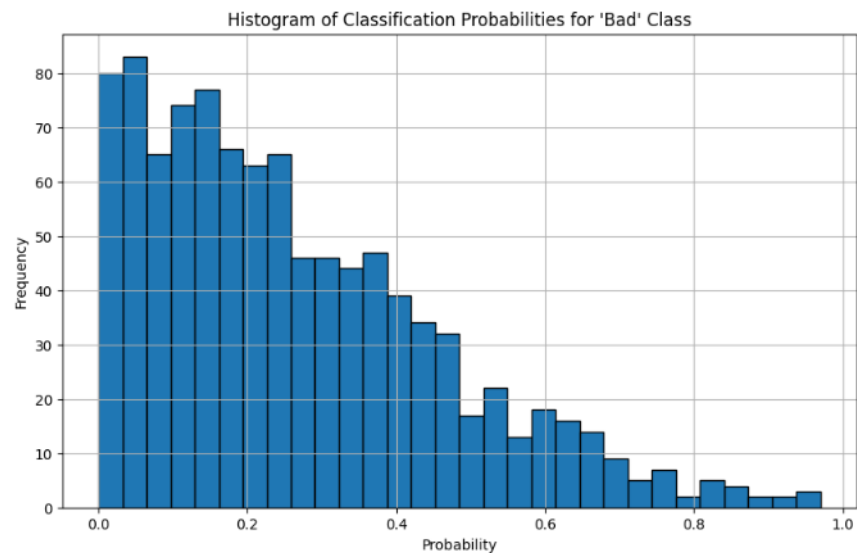
Examining the heatmap, it is apparent that a majority of superpixels positively influence the 'bad' classification, particularly those on the building's roof. Interestingly, the sky exhibits both positive and negative influences on the classification. The only superpixel contributing negatively (to the 'good' class) is the hood of the black car. The Coefficient of Variation, calculated from the weights, is equal to 1.73 for this image, indicating moderate variation among superpixel weights. This suggests that most superpixels contribute relatively equally to the overall explanation.

For the probabilistic classification, a lower, middle and upper threshold has to be chosen. In this research, the middle threshold is taken for -0.22, which as described in the face validity analysis is the average utility score of all the street view images analyzed. The upper and lower thresholds are 2 utility score points lower and higher. These thresholds are used for each case study.

- Lower threshold = - 2.22

- Middle threshold = - 0.22
- Upper threshold = 1.78

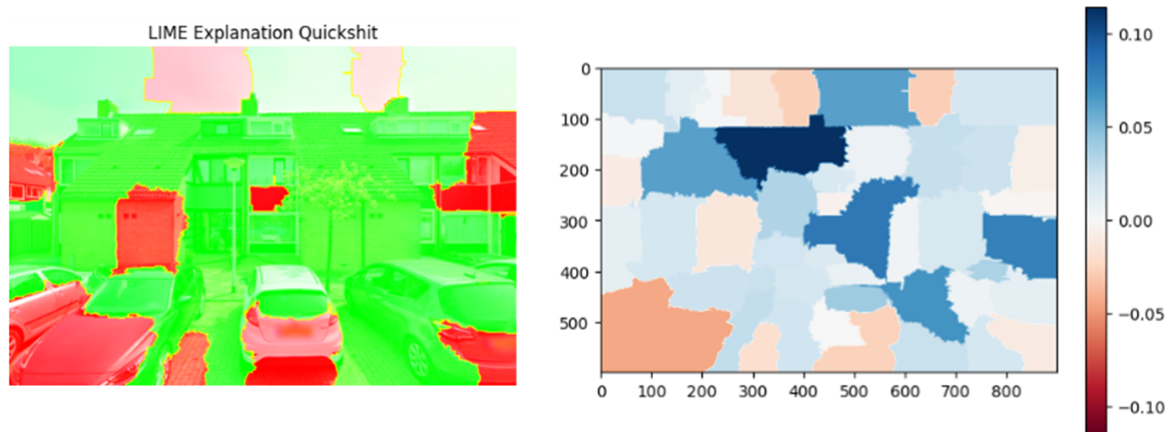
In terms of the segmentation algorithm and all other parameters for the LIME explanations, they are equal to the ones from the deterministic classification. First, the distribution of the probabilities for the 'bad' classification is shown to analyse the behaviour. The figure with probabilities for the classification 'good' is logically the mirrored figure of this figure. Figure 7.9 shows the distributions of the probabilities.



**Figure 7.9:** Distribution of the classification 'bad' probabilities in 1000 runs for the street view image of case study 1 with the probabilistic classification.

It can be seen that for the 'bad' classification the probabilities are mostly below 0.5. The BCR is equal to 0.133, thus not perfectly distributed between 'good' and 'bad' classifications. Important to note in this probabilistic classification algorithm is that the BCR is less important than in the deterministic case because the classification is scored with a probability that is more uniformly distributed than the 100 or 0 % deterministic case. The Probability Distribution Uniformity Metric allows for this analysis, and the distribution of these probabilities is skewed towards the left, with most samples present in the interval between 0 and 0,3 probability. This means that the probabilities for the classifications are not nicely distributed and thus not capture all the information for classification switches perfectly. Therefore, the LIME explanation is mostly fitted around this low probability of the classification 'bad' and is deemed as not a perfect explanation. Note here that a classification is deemed 'bad' if the probability is higher than 0.5 for this classification, and 'good' if otherwise.

The figure with all superpixel effects and the heatmap is shown in figure 7.10.



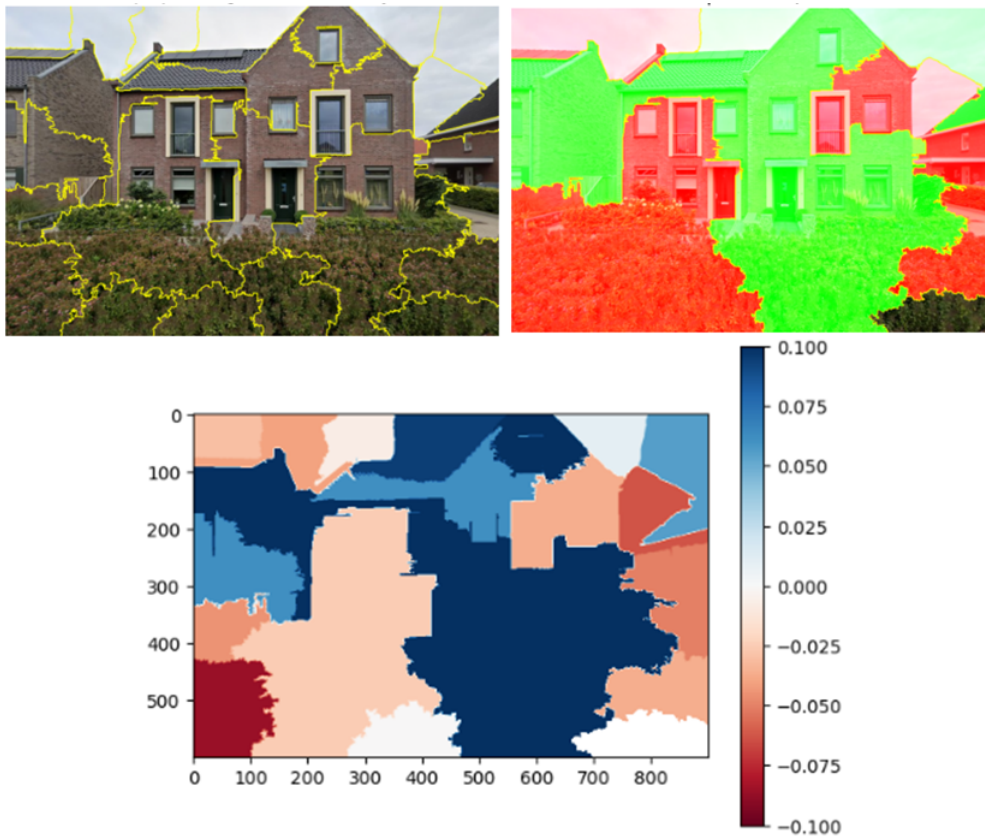
**Figure 7.10:** LIME explanation with all affecting superpixels and the heatmap for the street view image of case study 1.

As can be seen, an illogical explanation is given. If we compare this to figure 7.6, it can be seen that equal explanations are shown. The coefficient of variation is equal to 1.77, almost equal to the value with the other classification. Also, the pattern in the heatmap shows a resemblance to the other classification. Again, it can be concluded that LIME does not provide logical explanations to the user in this case.

6 other case studies are done to analyze the performance of the LIME explanations. A heatmap and plot with all influencing superpixels are shown for each case. The results concluded while the more in-depth analysis results can be found per case in Appendix J.

## 7.2. Case study 2

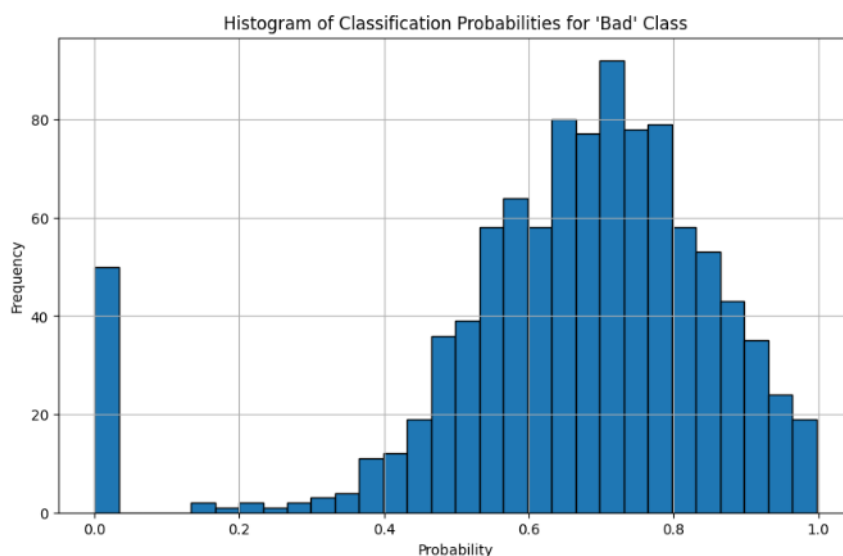
For case study 2, the original street view has a utility score of 0.38 and the classification threshold is 1.18 for the deterministic classification. Therefore, the LIME explanation concerns the classification of 'bad' for this image.



**Figure 7.11:** Case study 2: original street view image with segmentation, plot of all affecting superpixels and heatmap of superpixel weights in the deterministic classification in the probabilistic classification.

The positive weights thus contribute to the classification of 'bad'. In the analysis, it can be seen that parts of the building, as well as the green bush, contribute positively, as well as negatively to the classification of 'bad'. No clear semantic meaningful explanations can be concluded from this case study. Lime does not provide logical explanations to the modeller with the deterministic classification.

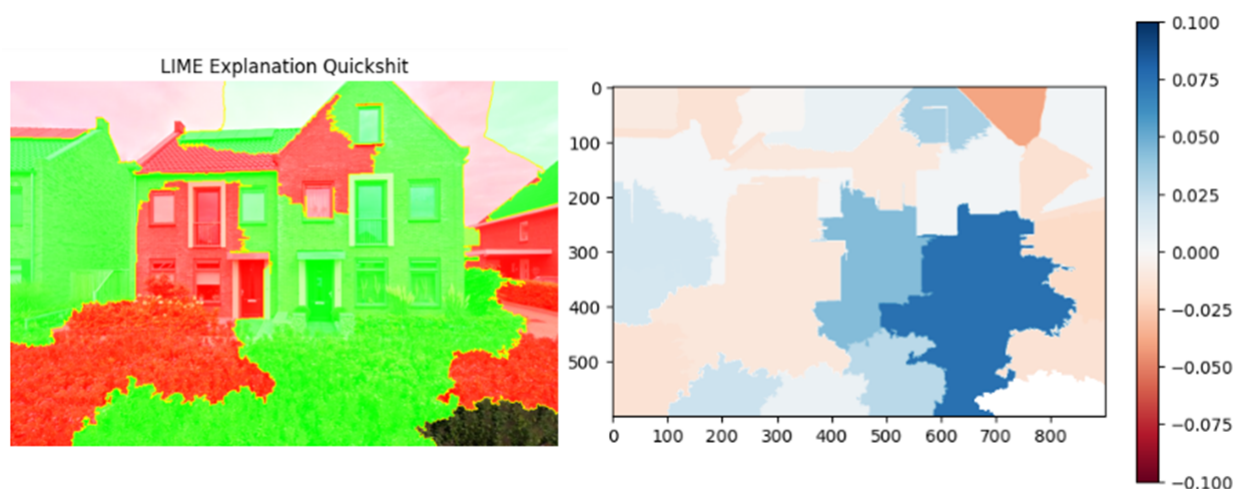
For the probabilistic classification, the distribution of the probabilities for the 'bad' classification is shown in Figure 7.12.



**Figure 7.12:** Case study 2: Distribution of the classification 'bad' probabilities in 1000 runs in the probabilistic classification

. It can be seen that for the 'bad' classification the probabilities are mostly above 0.5. The BCR is equal to 0.86, portraying not a perfect distribution of the bad and good classes. The perturbed images still have equal distribution as in the deterministic classification with a mean of 1.19. We can see that the distribution for the classification probabilities is skewed towards one side, therefore this LIME explanation is not perfect in quality.

The figure with all superpixel effects and the heatmap is shown in figure 7.13.



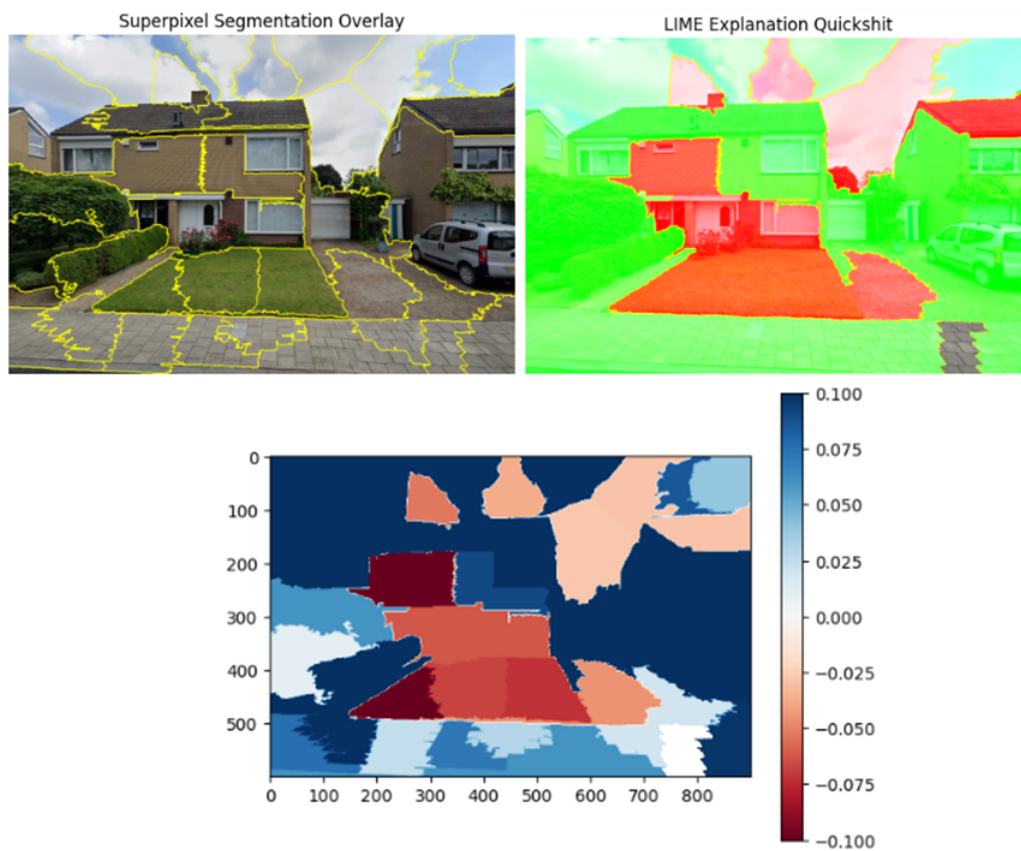
**Figure 7.13:** Case study 2: LIME explanation with all affecting superpixels and the heatmap in the probabilistic classification

As can be seen, an illogical explanation is given. If we compare this to figure 7.11, it can be seen that equal explanations are shown. The coefficient of variation is equal to 7.01. The pattern of the heatmaps (what contributes negatively and positively) is alike as well, but in the deterministic classification, the absolute values of the weights were higher than for the probabilistic classification. Again, it can be concluded that LIME does not provide logical explanations to the user in this case.



### 7.3. Case study 3

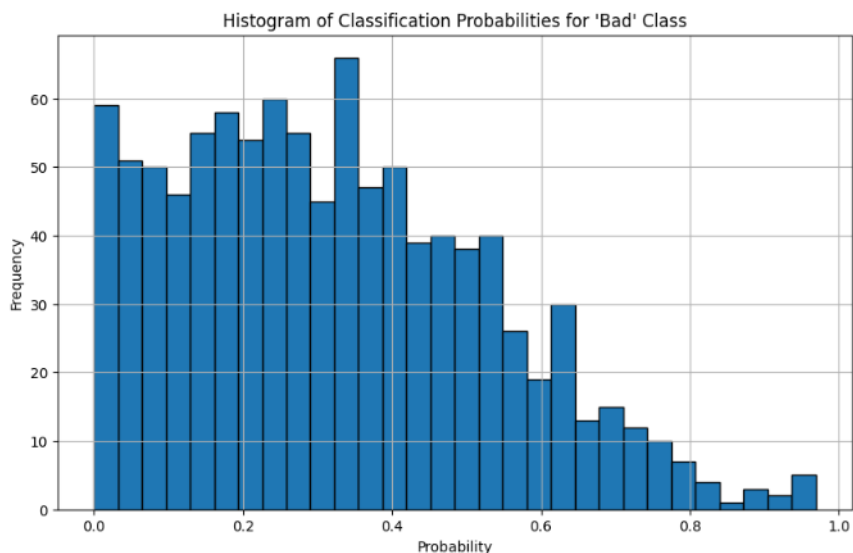
For case study 3, the original street view has a utility score of -0.11 and the classification threshold is 0.42 for the deterministic classification. Therefore, the LIME explanation concerns the classification of 'bad' for this image.



**Figure 7.14:** Case study 3: original street view image with segmentation, plot of all affecting superpixels and heatmap of superpixel weights in the deterministic classification.

The positive weights contribute to the classification of 'bad'. In the analysis, more semantic meaningful explanations can be seen. For example, the car seems to have a negative positive influence on the classification of 'bad', while a green lawn has a negative influence on the classification of 'bad', and thus contributes to the classification of 'good'. Also, the green bushes have a positive effect on the classification of 'bad', as well as the paved sidewalks and driveway.

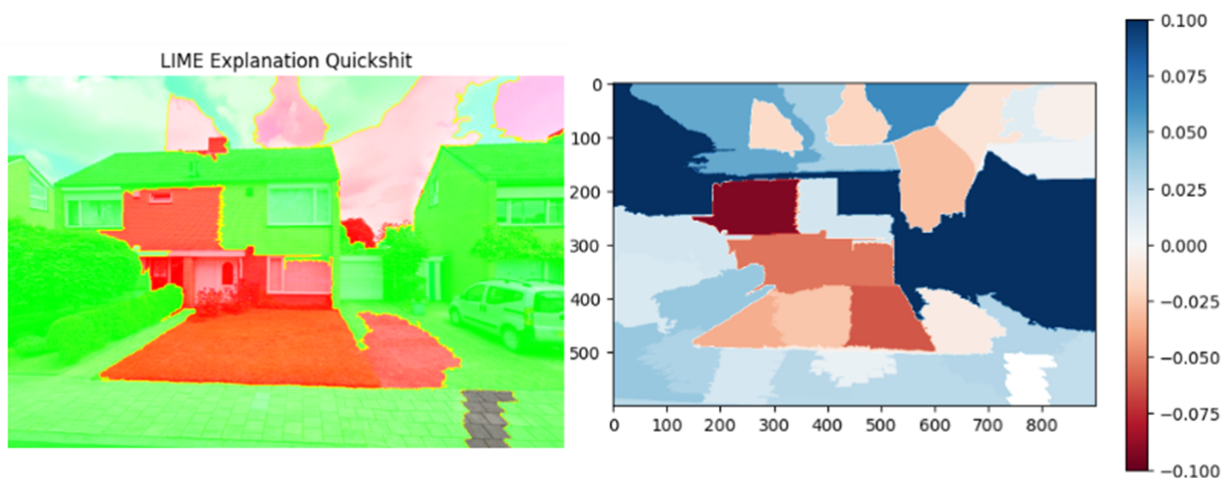
For the probabilistic classification, the distribution of the probabilities for the 'bad' classification is shown in Figure 7.15.



**Figure 7.15:** Case study 3: Distribution of the classification 'bad' probabilities in 1000 runs in the probabilistic classification

. It can be seen that for the 'bad' classification the probabilities are mostly below 0.5. The BCR is equal to 0.21, portraying not a perfect distribution of the bad and good classes. The perturbed images still have equal distribution as in the deterministic classification with a mean of 0.37. We can see that the distribution for the classification probabilities is skewed towards one side, therefore this LIME explanation is not perfect in quality.

The figure with all superpixel effects and the heatmap is shown in figure 7.16.

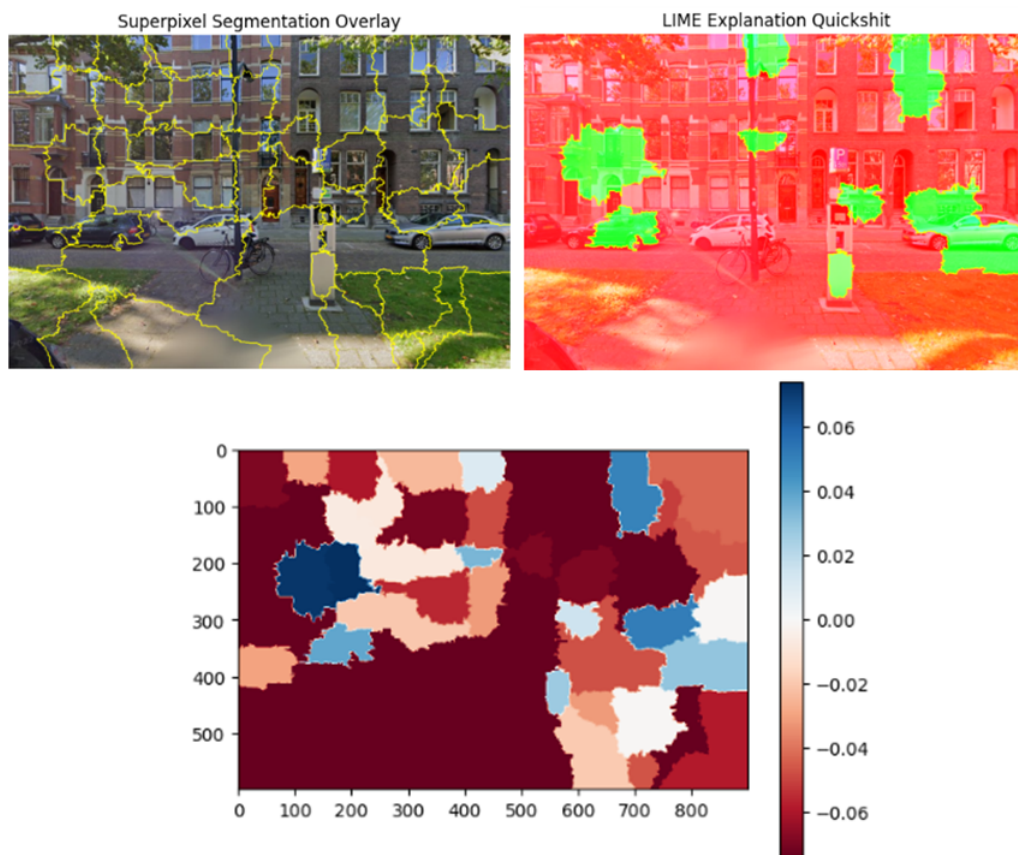


**Figure 7.16:** Case study 3: LIME explanation with all affecting superpixels and the heatmap in the probabilistic classification

As can be seen, a bit more logical explanation is given as we can see the direct influence of the porch on the classification. If we compare this to figure 7.14, it can be seen that equal explanations are shown. The coefficient of variation is equal to 4.11. The pattern of the heatmaps (what contributes negatively and positively) is almost alike as well. Again, it can be concluded that LIME does not provide logical explanations to the user in this case and that the different classification algorithms do not produce different results.

## 7.4. Case study 4

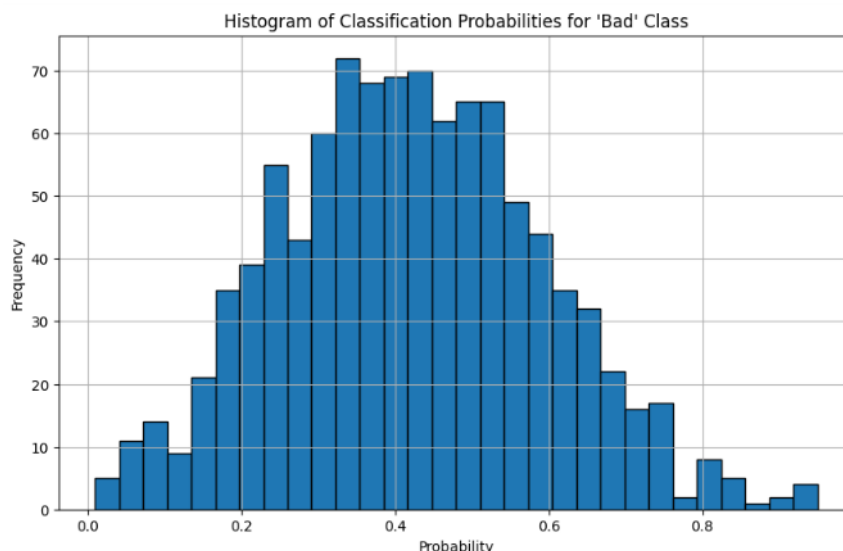
For case study 4, the original street view has a utility score of 0.11 and the classification threshold is 0.62 in the deterministic classification. Therefore, the LIME explanation concerns the classification of 'bad' for this image.



**Figure 7.17:** Case study 4: original street view image with segmentation, plot of all affecting superpixels and heatmap of superpixel weights in the deterministic classification.

The figure shows that most parts of the image contribute negatively to the classification 'bad', thus contributing to the 'good' classification. The only two clear features that contribute negatively to the classification are the greenery and the road, while for the cars and the home, their influence is positive as well as negative on the classification.

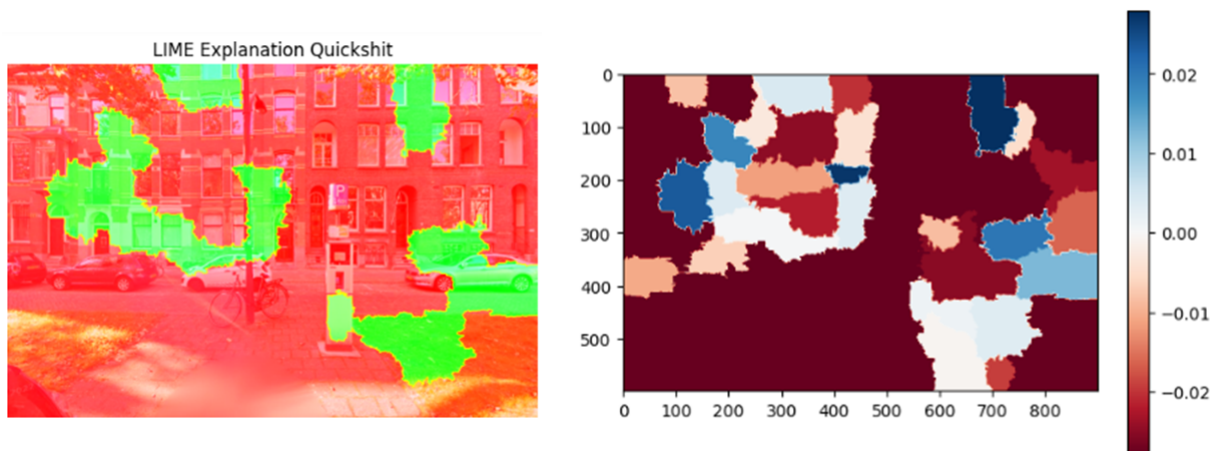
For the probabilistic classification, the distribution of the probabilities for the 'bad' classification is shown in Figure 7.18.



**Figure 7.18:** Case study 4: Distribution of the classification 'bad' probabilities in 1000 runs in the probabilistic classification

. It can be seen that for the 'bad' classification the probabilities are skewed a bit to the left so there are more 'good' predictions, but the distribution is more centred towards 0.5. The BCR is equal to 0.33, portraying not a perfect distribution of the bad and good classes. The perturbed images still have equal distribution as in the deterministic classification with a mean of 0.62.

The figure with all superpixel effects and the heatmap is shown in figure 7.19.



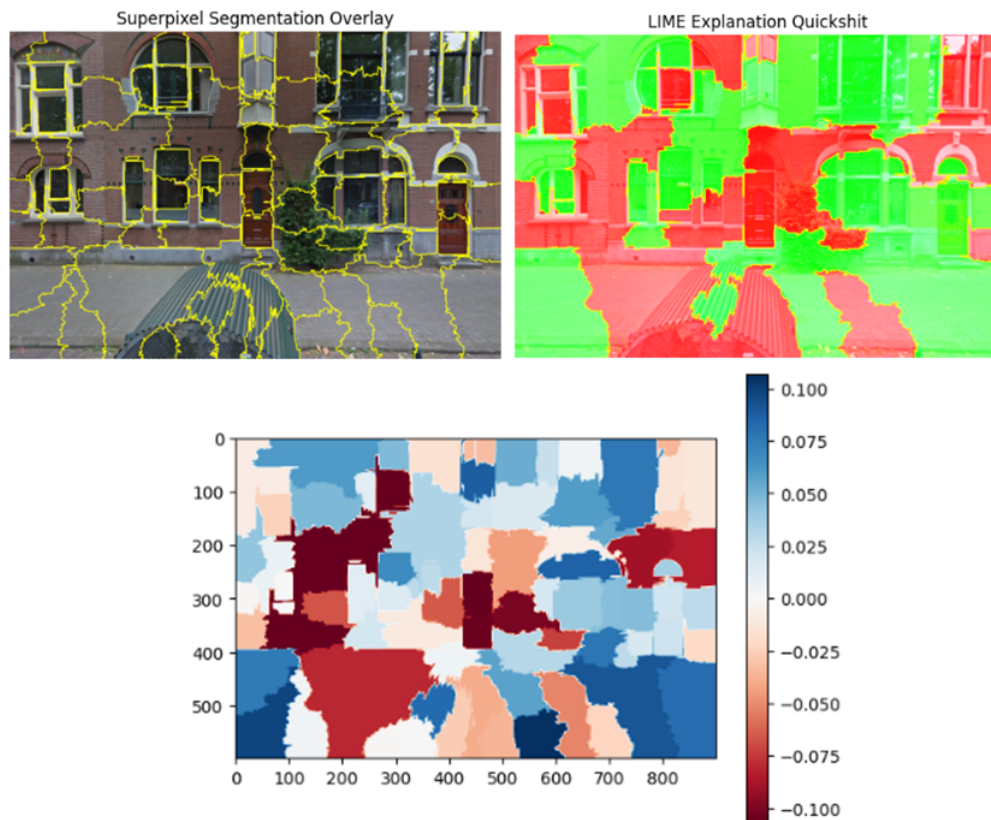
**Figure 7.19:** Case study 4: LIME explanation with all affecting superpixels and the heatmap in the probabilistic classification

As can be seen, the explanation is illogical, as no clear meaning can be drawn from it. If we compare this to figure 7.17, it can be seen that somewhat equal explanations are shown. The coefficient of variation is equal to -1.51. The pattern of the heatmaps (what contributes negatively and positively) is almost alike as well, showing mostly negative influencing images. Again, it can be concluded that LIME does not provide logical explanations to the user in this case and that the different classification algorithms do not produce different results.

### 7.5. Case study 5

For case study 5, the original street view has a utility score of -1.13 and the classification threshold is 0.28 for the deterministic classification. Therefore, the LIME explanation concerns the classification of

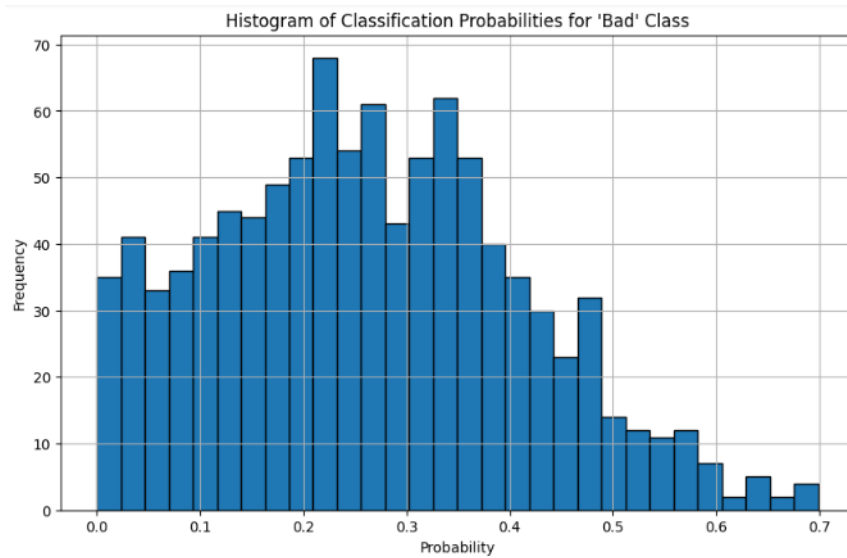
'bad' for this image.



**Figure 7.20:** Case study 5: original street view image with segmentation, plot of all affecting superpixels and heatmap of superpixel weights in the deterministic classification.

The figure itself already does not contain a lot of identifiable features, only the homes, the sidewalk, the bike storage and the bush could be identifiable objects. In the LIME explanation, none of these features affect only negatively or positively to the classification. Therefore, no clear logical conclusions can be drawn from this LIME explanation.

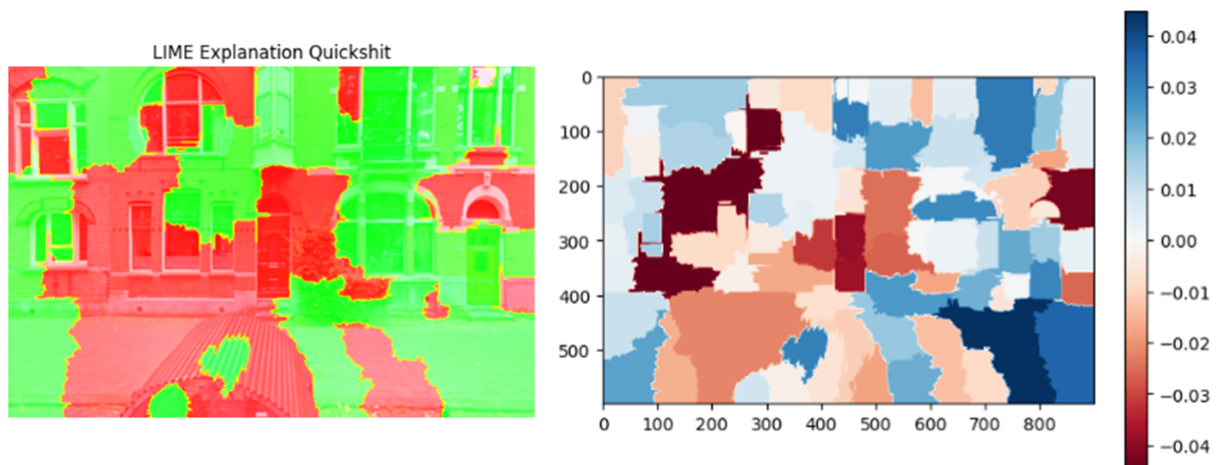
For the probabilistic classification, the distribution of the probabilities for the 'bad' classification is shown in Figure 7.21.



**Figure 7.21:** Case study 5: Distribution of the classification 'bad' probabilities in 1000 runs in the probabilistic classification.

. It can be seen that for the 'bad' classification the probabilities are mostly below 0.5. The BCR is equal to 0.06, portraying not a perfect distribution of the bad and good classes. The perturbed images still have a distribution as in the deterministic classification with a mean of 0.28. We can see that the distribution for the classification probabilities is skewed towards one side, therefore this LIME explanation is not perfect in quality.

The figure with all superpixel effects and the heatmap is shown in figure 7.22.



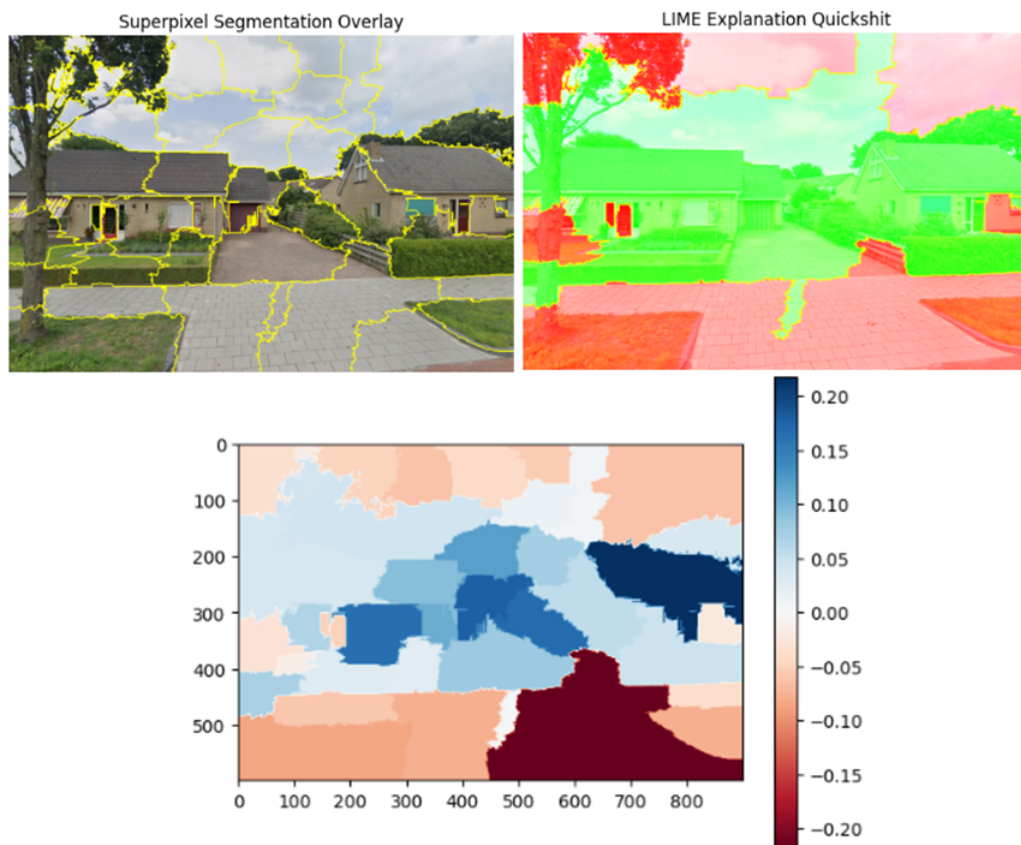
**Figure 7.22:** Case study 5: LIME explanation with all affecting superpixels and the heatmap in the probabilistic classification

As can be seen, the explanation is again illogical for the user. If we compare this to figure 7.20, it can be seen that equal explanations are shown. The coefficient of variation is equal to -24.89. The pattern of the heatmaps (what contributes negatively and positively) is almost alike as well. Again, it can be concluded that LIME does not provide logical explanations to the user in this case and that the different classification algorithms do not produce different results.

## 7.6. Case study 6

For case study 6, the original street view has a utility score of 1.15 and the classification threshold is 1.10 for the deterministic classification. Therefore, the LIME explanation concerns the classification of

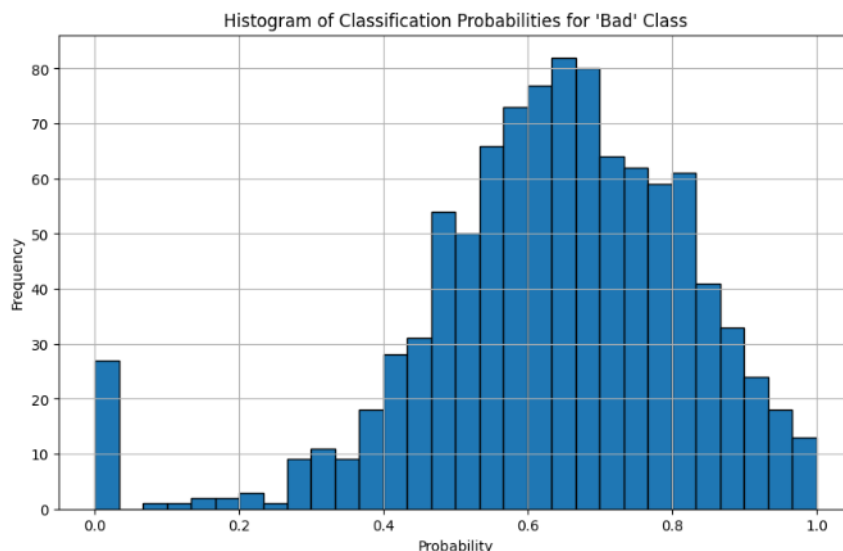
'good' for this image.



**Figure 7.23:** Case study 6: original street view image with segmentation, plot of all affecting superpixels and heatmap of superpixel weights in the deterministic classification.

The coefficient of variation is very high, 7.96, resulting in a diverse range of weights. This can also be seen in the heatmap. Next to this, both the heatmap and the 'all affecting superpixels' plot shows a clear distinction in this image. The upper and lower sections of the image contribute negatively to the classification. The influence of the tree is unclear, as it contributes positively and negatively to the classification. The presence of the sidewalks and the greenery near the sidewalk tend to have a negative influence, while the homes with their green gardens tend to have a positive influence on the 'good' classification.

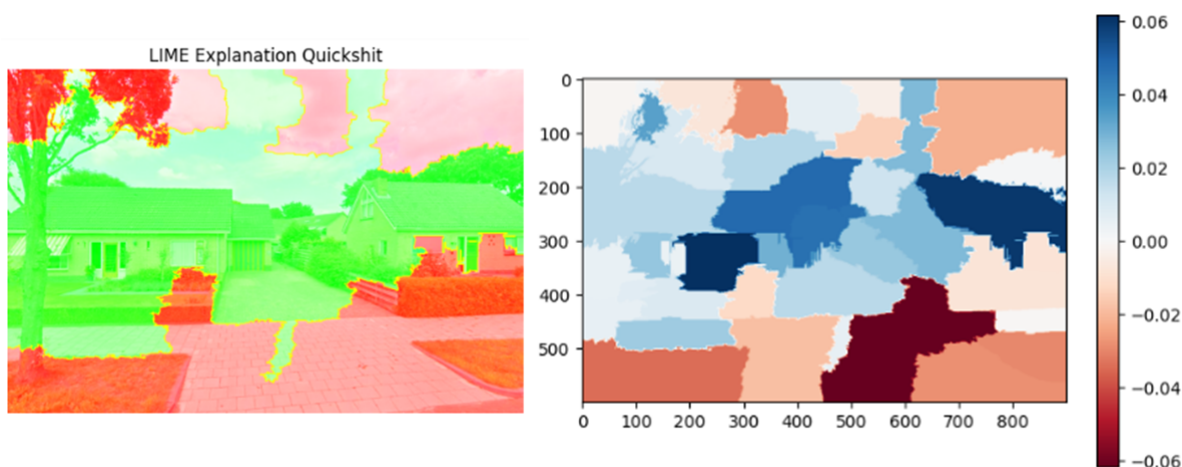
For the probabilistic classification, the distribution of the probabilities for the 'bad' classification is shown in Figure 7.24.



**Figure 7.24:** Case study 6: Distribution of the classification 'bad' probabilities in 1000 runs in the probabilistic classification

. It can be seen that for the 'bad' classification the probabilities are mostly above 0.5. The BCR is equal to 0.8, portraying not a perfect distribution of the bad and good classes. The perturbed images still have a distribution as in the deterministic classification with a mean of 1.11. We can see that the distribution for the classification probabilities is skewed towards one side, therefore this LIME explanation is not perfect in quality.

The figure with all superpixel effects and the heatmap is shown in figure 7.25.



**Figure 7.25:** Case study 6: LIME explanation with all affecting superpixels and the heatmap in the probabilistic classification

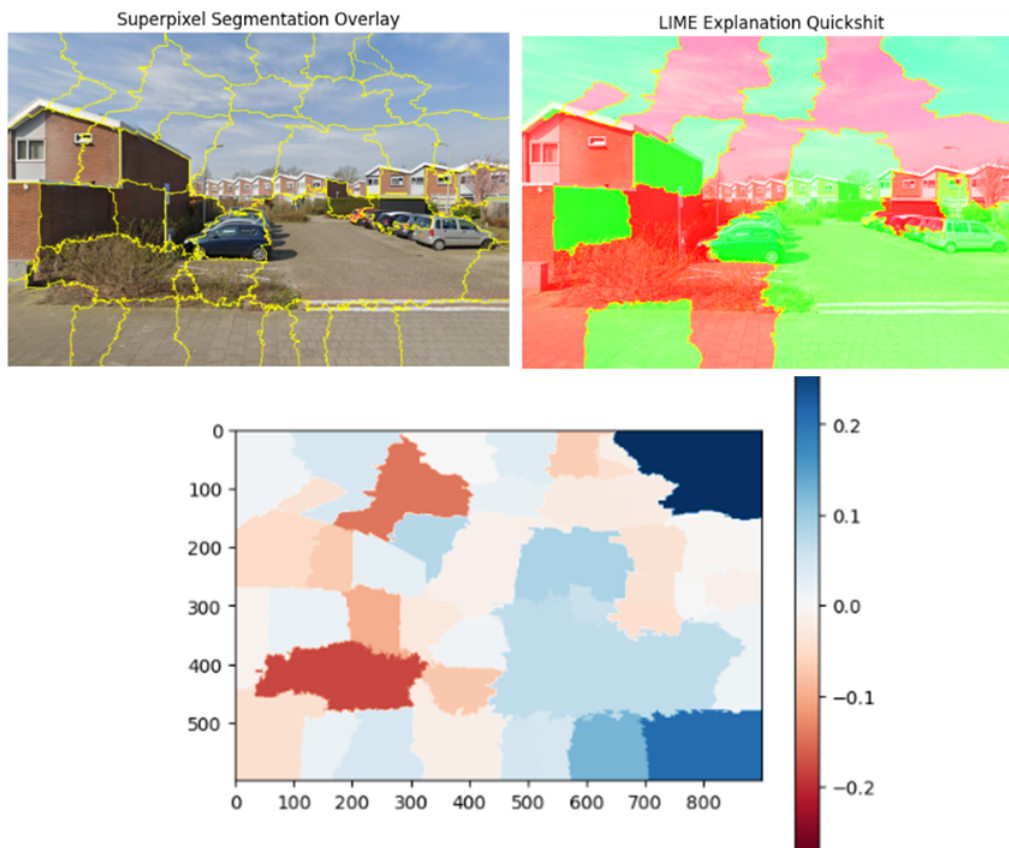
As can be seen, the explanation is again illogical for the user. If we compare this to figure 7.23, it can be seen that equal explanations are shown. The coefficient of variation is equal to 3.8. The pattern of the heatmaps (what contributes negatively and positively) is almost alike as well. Again, it can be concluded that LIME does not provide logical explanations to the user in this case and that the different classification algorithms do not produce different results.

### 7.7. Case study 7

For case study 7, the original street view has a utility score of -0.26 and the classification threshold is 1.47 for the deterministic classification. Therefore, the LIME explanation concerns the classification of



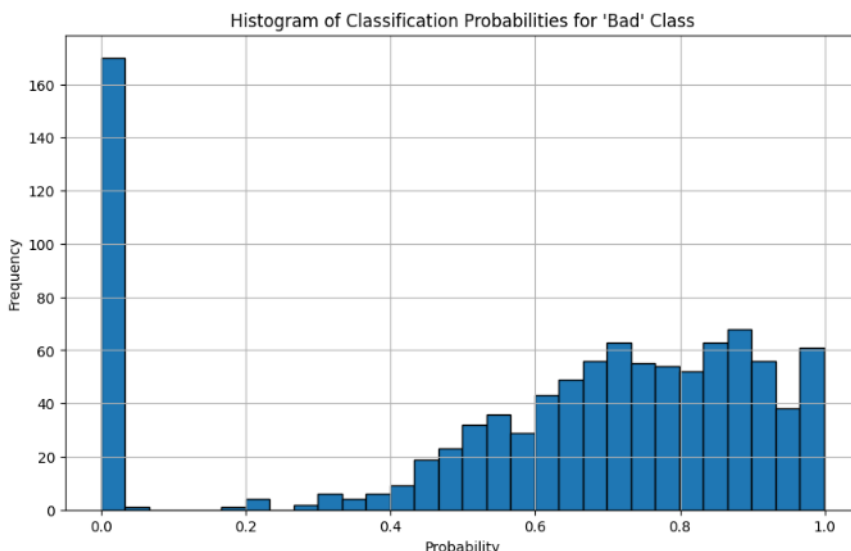
'bad' for this image.



**Figure 7.26:** Case study 7: original street view image with segmentation, plot of all affecting superpixels and heatmap of superpixel weights in the deterministic classification.

The LIME explanation does not provide clear objects influencing the classification in only one way, except for the cars, which contribute positively to the classification of 'bad'. For the other features, no clear conclusion can be drawn.

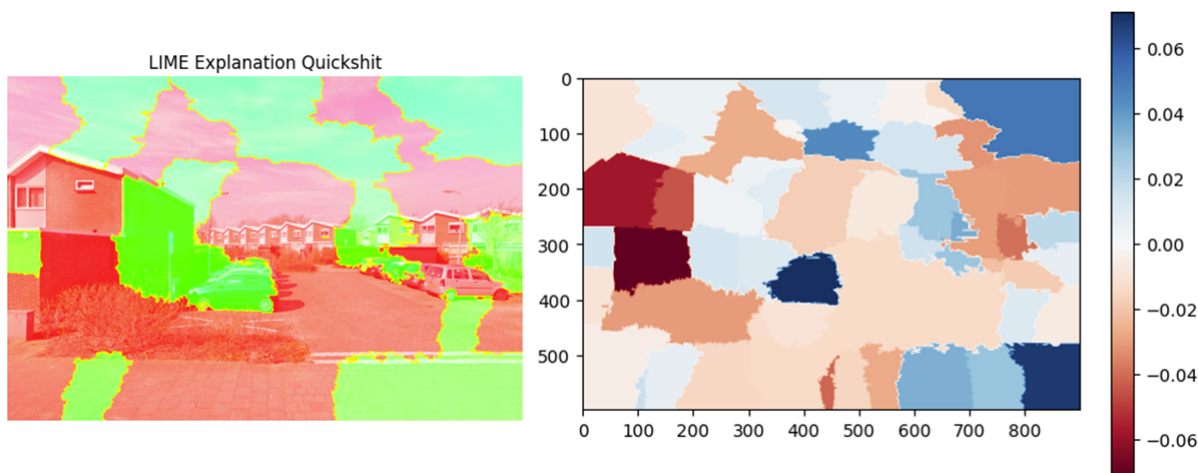
For the probabilistic classification, the distribution of the probabilities for the 'bad' classification is shown in Figure 7.27.



**Figure 7.27:** Case study 7: Distribution of the classification 'bad' probabilities in 1000 runs in the probabilistic classification

. It can be seen that for the 'bad' classification the probabilities are mostly above 0.5, while the 0% probability is seen the most, which is interesting. In the sampling process, it has created images which are 100 % sure 'good'. The BCR however is 0.75, showing the high amount of 'bad' classifications. The perturbed images still have a distribution as in the deterministic classification with a mean of 1.37.

The figure with all superpixel effects and the heatmap is shown in figure 7.28.



**Figure 7.28:** Case study 7: LIME explanation with all affecting superpixels and the heatmap in the probabilistic classification

As can be seen, the explanation is again illogical for the user. If we compare this to figure 7.26, it can be seen that the explanations are somewhat equal. Some superpixels have switched from positive to negative or vice versa, but most remained the same contribution. This behaviour can also be seen in the heatmap, where these differences in influence can more clearly be analyzed. The superpixels which changed do not show clear semantic meaning. The coefficient of variation is equal to -40.91, which is high. This is following the more intense superpixel contributions which can be seen in the heatmap. Again, it can be concluded that LIME does not provide logical explanations to the user in this case and that the different classification algorithms do provide some different explanations, but the overall behaviour of the explanation remains the same.

## 7.8. Application of LIME conclusion

The application of LIME aims to enhance the explainability and interpretability of the model. Next to this, it also aims to provide human-understandable explanations about the complex model's decision behaviour on a certain output. In the case studies, LIME provides explanations about the decision behaviour of the model for each street view with the image's unique classification. However, these explanations are unfortunately not humanly understandable, they are not logical to the user. Therefore, the user does not increase their knowledge about how the complex model makes decisions on the utility-based liveability score, and therefore LIME does not increase the explainability and interpretability of the model for the user.

From the seven use cases with both classifications, it is evident that in LIME's current state of methodology and implementation, LIME does not provide additional human understandable explainability to the modeller regarding the decision behaviour of the complex model. In most case studies, and with both classification algorithms, the LIME explanation results are unclear and illogical to the modeller. Nonetheless, in case studies 3 and 6, the LIME explanation does seem to provide logical insights. This logical insight is gained from LIME explaining some parts of the street view images which have semantic meaning. This results in the problem of this LIME implementation that the semantic meaning in the explanations is missing. The segmentation algorithms result in the street view image having superpixels that when some are combined, form a human-understandable object such as a car or tree. But in the explanations, most times these superpixels are not all equally influencing the classification, therefore resulting in objects that some parts contribute negatively and some positively to the classification, which is not humanly interpretable. From the perceived liveability literature, multiple street view objects were stated to influence the liveability, unfortunately, the explanations did not have semantic meaning representing any of these objects to analyze whether the same effect is seen in this research for these objects. This problem of semantic meaningful superpixel explanations has also been found in the other explanations next to the seven case studies provided here.

These current LIME explanations lead to a fragmented and confusing interpretation for human users. The primary reason for adopting XAI techniques was to avoid the biases introduced by prior object detection, which predetermines the elements for analysis and thus constrains the model's interpretability. By employing a post hoc analysis, the integrity of holistic human perception is maintained without introducing initial biases. However, to achieve semantically meaningful explanations, it is crucial to ensure that each object in the street view corresponds to a single, coherent superpixel. This alignment allows the explanations to be more intuitive and understandable, reflecting the true influence of individual objects in the street views. Without this coherence, as seen in the cases, the explanations lack the necessary semantic clarity, undermining the objective of making the model's decisions comprehensible and useful for practical applications.

However, the metrics show that the explanations are scientifically correct. The Jaccard and DICE values demonstrate that the segmentations are adequate or good, and the LIME metrics show a good BCR for each case study with the deterministic classification. This indicates that the LIME explanations are trained on sufficient data and thus have significant value. In the probabilistic classification, the BCRs tend to be not good (not close to 0.5), but the Probability Distribution Uniformity Metric shows different behaviour as for some explanations the distribution of the probabilities is more uniform than in others, while no distribution seems to perfectly be uniformly distributed. Therefore, not all explanations for the probabilistic classification are scientifically perfect. The coefficient of variance is adequate in all cases, with some cases even being very good (higher than 5 in absolute value). Therefore, in the deterministic classification, the LIME explanations are scientifically grounded by the metrics used and do provide insights into the model's decision behaviour, while in the probabilistic, the explanations are good but do not always have a perfect PDUM, scientifically less valid than the explanations from the deterministic classification.

The classification algorithms both provide almost equal explanations in terms of the negative or positive influence of superpixels. In absolute value, the influence of superpixels can sometimes differ between classification techniques. But the explanation they provide is almost equal between the classifications. The BCR for the deterministic classification is good, while in the probabilistic classification, this is not

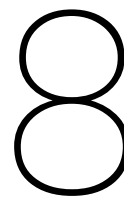
the case. For a good explanation, a more uniform distribution of the classification probabilities is wished, which is not seen in the 7 cases discussed in this research. For user-friendly explanations, this probabilistic classification does provide benefits as it is easier to explain and compare with other explanations, but since the BCR value is not perfect here, no clear decision can be made for which classification technique works better, however, due to the equal results from both classification algorithms, it suggests that the model performs consistently, also regarding different classification models, and thus is robust in explanations.

However, it is important to note that LIME attempts to explain a complex model. For LIME to provide logical explanations, it is assumed that the complex model's decision behaviour is complex but meaningful. This results in that there is the possibility that the complex model itself has inherently illogical decision behaviour, suggesting that it should not be used for policy advice and requires further improvement and that LIME shows this illogical behaviour of the model. But, this conclusion is not grounded as there is no proof for it, it is only a possible explanation for the bad human-understandable LIME explanation.

The face validity analysis indicated that the presence of greenery is a clear beneficial factor for a high utility score. However, the LIME explanations do not consistently support this conclusion, as greenery is not always positively correlated with a 'good' classification, whereas the same holds for cars and water presence. To answer the sixth sub-question, the current LIME explanations do not provide additional insights beyond the face validity analysis.

This leads to two possible conclusions: First, the LIME explanations do not increase the explainability for the decision behaviour of the complex model, and also do not provide human-understandable LIME explanations. Due to the LIME explanation showcasing low semantic meaning, the user can not directly analyze which object or relations between objects influence the classification of the street view image. The superpixel creation does not divide the street view image into human-understandable objects. Additionally in the explanation of LIME, these different superpixels for a single object contribute much differently to the classification of good or bad (positively and negatively). This leads to that currently, this computer vision-enriched discrete choice model with LIME explanations can not be used for creating policy advice. As for policy advice, the explanations of how the objects in street view images contribute to urban liveability are not clear, and thus no policy advice can be given. Second, there is the possibility that the complex model itself has inherently illogical decision behaviour, providing LIME no clear basis to create logical explanations.

Another drawback to the application of LIME outside the technical working of the explanation is the computational cost of one explanation. The process of creating and judging the correct segmentation process, creating the ground truth image, analysing the perturbed image scores and creating the LIME explanation are four time-consuming steps. Creating one explanation in the current pipeline, thus executing each step correctly, can take around 1 hour in total. The current pipeline process is not optimised, but even if optimised this process will take a long time to produce one explanation for one street view image. This is a disadvantage of using LIME in this application for policy advice. The CUDA version used is 12.3 and the CPU is 12th Gen Intel(R) Core(TM) i7-12700H 2.30 GHz.



# Discussion

The results of the previous chapter indicate that the application of LIME in order to enhance the interpretability of computer vision-enriched discrete choice models does not provide logical human understandable interpretations for users and also does not increase the explainability of the complex model. Several steps in the LIME application pipeline contribute to this result, each having limitations and implications.

## 8.1. Sampling Process and Utility Score Distribution

The utility score distribution in the sampling process can be further researched. Experimenting with different parameters or distributions beyond the Bernoulli distribution with a 0.5 parameter could be beneficial. Vermeire et al. [135] and Rashid et al. [110] state that in this current approach for images with LIME the neighbourhood of the explained sample is under-represented, resulting in under-representation of the true local behaviour of LIME around the input image. LIME works by creating a local approximation of the model's behaviour around a specific instance by generating perturbed versions of the original image and observing the model's responses to these perturbed images. If the perturbed samples do not adequately represent the original image's local neighbourhood, the resulting explanations may not accurately reflect the model's decision-making process for the actual image. Without properly representing the original image's neighbourhood, LIME's linear surrogate model may not capture the complex, non-linear relationships in the model's decision process. This loss of local fidelity results in explanations that do not align well with the model's actual behaviour for the input image. Further research into this under-representation could improve the LIME image explanations.

## 8.2. Complexities of Street View Images

Street view images are complex images in the context of urban research for several reasons, as was stated by Law et al. [66]. Firstly, they encompass a vast array of visual elements. Each of these elements can have different textures, colours, shapes, and sizes, adding to the overall visual complexity of the image. Secondly, street-view images capture dynamic and diverse urban environments. As seen in the face validity analysis, street view images are very diverse and do not have a standard of what is contained in them. This makes it difficult to achieve perfect segmentation with standard algorithms that are applicable currently for LIME. Each image may require a different segmentation approach, with varying parameters, leading to segmentations that are not always semantically meaningful. As seen in Chapter 7, the segmentation can segment important objects, but often in such a way that a single object is made up of multiple superpixels. This results in LIME explanations being illogical, as certain superpixels of a single object affect the classification both positively and negatively. For instance, one part of a tree might contribute to the classification 'bad' while another part of the same tree contributes to the classification 'good'. The evaluation of the quality of these segmentations can either be done subjectively by the user and/or scientifically via a ground truth image and additional metrics. Since there are multiple standard options for LIME's segmentation algorithms (which all can also use different parameter values), examining all options for each input image is not effective. This

research thus assumes some segmentation algorithms to be best for analysis, judged by the modeller. Additionally, the full parameter ranges for these segmentation algorithms are also not analyzed. Due to the importance of the creation of correct superpixels (Knab, Marton, and Bartelt [65], Schallner et al. [118]), this creates a high limitation, as it is not currently possible to analyze all possible algorithms with perfect parameter finetuning. Object detection for street view is currently available, but not yet applicable to superpixel segmentation algorithms for LIME. Further analysis can be done by using street view trained object detection models for superpixel creation, like those in the research by Cordts et al. [22].

### 8.3. Need for prior object detection

Prior object detection is not used to prevent introducing biases and pre-determining how features are measured. However, prior object detection for superpixels is needed because it ensures that each superpixel corresponds to a single human-understandable object within the street view image, addressing the issue of illogical explanations from the previous chapter. Applying object detection solely for superpixel creation ensures that superpixels correspond to complete objects, maintaining the integrity of the holistic view of human perception. This approach introduces a dependency on the accuracy and quality of the object detection algorithm, acknowledging the inherent bias it introduces, affecting how well the segmentation aligns with the actual objects in the image. Despite introducing this bias, the LIME explanation remains beneficial as it can still provide local explanations, capture complex relationships in street views, and offer user-friendly explanations for possible policy advice. The need for object detection for superpixel creation is evident to reduce computational time and produce more humanly understandable explanations. However, understanding the workings of the object detection model is crucial to managing the introduced bias effectively.

### 8.4. Subjective nature of LIME

The analysis of LIME outcomes is inherently subjective because, while LIME provides objective explanations, the quality of the explanation is judged on how understandable the explanations are. Metrics like the Binary Classification Ratio, the Probability Distribution Uniformity Metric, and the Coefficient of Variance can objectively evaluate LIME's explanations. Currently, there is limited research on LIME metrics, meaning that the metrics used in this research may not be the best metrics to objectively evaluate LIME explanations on their quality.

### 8.5. Classification problem

In the current implementation of LIME, converting a regression model (the CVDCM) into a classification model is necessary for generating explanations in the current methodology (in theory, LIME for images can work for regression but the current Python implementation does not allow it yet). This requirement significantly influences how LIME explanations are produced, with the need to categorize outcomes into binary classes, such as good or bad. In this research, both deterministic and probabilistic classification algorithms are applied to facilitate this binary categorization. For policy advice purposes, field experts can set appropriate thresholds to define what constitutes good and bad utility-based liveability scores, enabling LIME to evaluate these classifications effectively. However, the deterministic classification presents challenges. Due to the nature of perturbed image scores and the requirement for the Binary Classification Ratio (BCR) to be around 0.5, employing a single, consistent threshold across all images is impractical. Each image ends up needing a unique threshold, leading to varying definitions of "good" and "bad" for different cases. This approach risks oversimplifying the nuanced information captured by the original regression model, potentially losing important details that influence the utility-based liveability score in the process. In contrast, probabilistic classification can accommodate these variations more effectively. By assigning probabilities to each class, it retains more of the nuanced information present in the regression model. This approach aligns better with policy advice tools, allowing for a more flexible and detailed analysis that can adapt to the specific requirements of different contexts. However, it's crucial to note that both methods introduce a level of abstraction that may obscure some of the granular insights provided by a pure regression model. The necessity to convert continuous outcomes into discrete classes can inherently lead to a loss of information, particularly in capturing the subtleties of the utility-based liveability score. Future updates to LIME packages in Python may address this limitation

by enabling direct application to image-based regression models, thereby preserving the full richness of the original data. Further research can thus be done on possible other implementations of LIME for images where regression models can be used. Moreover, if future research incorporates the use of object detection for superpixel creation, probabilistic classification is recommended. This method allows for easier comparison between images and provides more comprehensible explanations, as it can handle the variability and complexity of urban environments more effectively.

## 8.6. Time costs

The current LIME application process is time-consuming and complex, requiring careful segmentation, analysis of perturbed image distributions, and generation of explanations for each image. This complexity demands significant expertise and is not easily scalable for analyzing large volumes of images, as needed for municipal policy analysis. Although LIME explanations can enhance explainability, the computationally intensive pipeline makes large-scale street view image analysis impractical. To address this issue, introducing object detection segmented superpixels can simplify explanations and reduce computational time, as the same segmentation algorithm can be applied to each image. Additionally, deterministic and probabilistic classifications yield similar results, with probabilistic classification eliminating the need for a unique threshold. Implementing an optimized segmentation algorithm and probabilistic classification can significantly reduce these computational costs.

## 8.7. Comparison with Traditional Object Detection Models

In exploring the necessity of using XAI like LIME to enhance model explainability, it is crucial to compare this approach with traditional street view object detection models to evaluate the necessity of XAI usage. Traditional object detection models segment images into predefined objects and analyze their features to determine their contributions to the utility-based liveability score. This method offers simplicity and direct interpretability, correlating specific objects with utility scores and facilitating more straightforward insights. Training a model on detected objects and utility scores is less time-consuming than using LIME, providing clear, global explanations of the model's decision behaviour (e.g., cars influence X much, trees Y much). Assuming an object detection model for street views is used for the object detection model and to create superpixels in LIME, both models would share the same bias regarding with which features the utility score is measured. The simple object detection model efficiently explains what influences the model's decision behaviour based on these objects. However, this simplification might overlook complex interactions between various elements in street view images, potentially missing nuanced influences on the utility score and failing to incorporate the holistic view of human perception in street views. LIME can analyze the holistic view of the human urban perception. Moreover, traditional models cannot provide local explanations, limiting their ability to explain specific outcomes, which LIME does, making it useful for targeted policy advice for certain streets or neighbourhoods. This generates a better understanding of the model's decision behaviour, valuable for models used in policy advice. Therefore, using LIME over simple object detection models is justified because LIME offers a more comprehensive analysis with actionable insights, making it a more powerful tool for urban research and policy development.

### 8.7.1. Traditional object detection models with tabular LIME explanations

For these traditional object detection models to generate local explanations, LIME for tabular data can be applied. LIME can also provide local explanations for tabular data, detailing how each object contributes to the liveability score for specific images through feature importance. This approach allows for an understanding of the influence of different objects within an image, potentially capturing complex relationships between the objects and nuanced influences. The benefits of using LIME for tabular data with traditional object detection models include simplicity, clarity, and reduced computational costs. This method simplifies the explanation process by focusing on well-defined objects, providing feature importance for each object to utility scores. The reduced computational costs make it practical for municipal policy analysis. Additionally, this method reduces the need for classifying the model, as tabular LIME is implemented easily for regression. The local explanations of LIME for tabular data provide the ability to model the holistic view of human perception by analyzing each image and the influences of the objects present. Due to the suggested use of object detection for superpixel creation, the explanations for LIME images and LIME tabular become similar, as both explanations are human-

understandable based on the presence of these objects. However, a significant advantage of LIME for image analysis is the direct visual representation, making it immediately useful for policy advice by clearly illustrating what happens in the images. With LIME for tabular data, while we understand the objects' influence, the image must be shown alongside the tabular explanations to comprehend the context fully. This allows the possibility of still analyzing the holistic view, while also enabling the identification of more nuanced differences between the utility score images. In this research, due to the need for object detection-created superpixels and the complication of classifying the regression model, there is a possibility that using an object detection model trained on the utility score with tabular LIME explanations could be beneficial over the proposed image LIME methodology. However, complications with LIME images arose, and it cannot be guaranteed that similar issues will not occur with tabular LIME. Further research is necessary to validate the effectiveness and address potential issues with tabular LIME. Another disadvantage of possibly using tabular lime with the object detection model is that a machine learning model must be trained on the objects and the utility score first to create the traditional object detection model. Therefore, the quality of LIME explanations will depend on the accuracy and performance of this machine-learning model.

### 8.7.2. Other image XAI techniques

Several other XAI techniques can provide insights beyond what LIME (Local Interpretable Model-agnostic Explanations) offers. Among these, Grad-CAM (Gradient-weighted Class Activation Mapping) and SHAP (SHapley Additive exPlanations) are notable alternatives (Van der Velden et al. [132], Alicioglu and Sun [4]). Each of these methods differs from LIME in their approach to creating human-understandable explanations and enhancing the interpretability of complex models.

Grad-CAM provides an approach to interpreting model predictions by generating heatmaps that highlight important regions in an image. Unlike LIME, Grad-CAM uses the gradients flowing into the final convolutional layer of a neural network to produce a localization map. This map indicates which parts of the image are most relevant to the prediction. One significant advantage of Grad-CAM over LIME is its ability to maintain the integrity of the entire image without requiring segmentation into superpixels. This avoids the problem of illogical explanations that can arise when LIME's superpixels do not align with human-understandable objects. Grad-CAM produces a direct visualization of the model's focus areas, which can be more intuitive for users. However, there is no guarantee that the heatmaps created by Grad-CAM are semantically meaningful and thus applicable to policy advice, as they provide the user with a heatmap with important regions, not certain objects (Selvaraju et al. [119]).

SHAP provides a measure of feature importance by considering all possible combinations of features of an image. SHAP uses also superpixels as features but provides a different method of calculating the importance of superpixels. However, SHAP can also use other techniques for creating features in an image, such as working with pixels or with intermediate layers of the convolutional neural network of the complex model, but superpixels are the most used feature creation in images. SHAP provides a more consistent and theoretically sound explanation than LIME, which is less robust. The primary advantage of SHAP over LIME in this context is thus its robustness in providing consistent feature attributions, which can improve the reliability of the explanations, and different superpixel weight calculations. Additionally, SHAP does not require the conversion of regression models into classification models, allowing for the direct interpretation of regression outcomes (which theoretically LIME also can, but not yet easily applicable in LIME. Note here that there is no 100 % guarantee that SHAP or GRAD-CAM had this easy regression implementation for Python). However, SHAP also faces challenges. Its computational costs can be high. Also, its feature importance measurement on superpixels holds the drawback that these explanations are dependent on the quality of this superpixel creation again (Lundberg and Lee [83]).

Therefore, SHAP and GRAD-CAM are suggested for further research to analyze the applicability of these techniques to provide interpretability of the complex model and human-understandable explanations.



# 9

## Conclusion

The goals underlying this research were to analyze whether applying XAI techniques like LIME for computer vision-enriched discrete choice models with street view images enhances the explainability and interpretability of these models, while also providing human-understandable explanations which can be used for policy advice. The methodology consisted of applying the CVDCM model for a face validity analysis to the city of Rotterdam, applying LIME to these results and verifying the results with metrics for the segmentation results and the LIME explanations. The research indicates that LIME currently can not provide human-understandable interpretations, and its implementation does currently not provide benefits to the CVDCM model, nor does it increase the utility of these models for policymakers. This is due to major problems occurring during the superpixel creation and classification phases.

The primary problem with the current LIME implementation lies in the segmentation algorithm's quality. LIME relies heavily on effective image segmentation to create superpixels, which is particularly challenging with street view images due to their complexity, variety of objects, and the mutual difference between street view images. Ground truth analysis can provide scientifically valid segmentation choices, but this process is labour-intensive and can introduce inconsistencies in quality. Current segmentations often result in superpixels that segment a single object into multiple parts. Although these combined superpixels have semantic meaning, the explanations often indicate conflicting contributions from different parts of the same object, leading to interpretations that are not humanly understandable. As a result, while LIME provides information about the decision behaviour of the CVDCM, the explanations are frequently illogical to users due to the lack of semantically meaningful superpixels. Users cannot effectively interpret the model's decisions, meaning the current LIME implementation does not enhance explainability. Despite the scientific validity of LIME's explanations, they suggest that the decision behaviour of the CVDCM might be inherently complex and illogical, or cannot be captured in terms and concepts used by people. This makes it un-interpretable currently for possible use in policy advice.

The second problem is with the classification. The behaviour of perturbed images in the LIME sampling process results in sampled images that differ significantly from the original image, affecting the utility-based liveability score. This variability influences the classification threshold for the deterministic classification, leading to unique explanations for each image and complicating the comparison of results across different images. This is also due to the assumption made that the LIME explanations should have a good Binary Class Ratio (BCR). This complexity makes the application of these explanations for policy advice less useful. In this research, both a probabilistic and deterministic classification are analyzed, producing similar explanations and therefore providing robustness to the LIME explanations. While the LIME explanations with the deterministic classification are more scientifically valid (perfect BCR and good CoV (Coefficient of Variation)), the probabilistic explanations (worse BCR, not perfect PDUM (Probability Distribution Uniformity Metric), and good CoV) provide a better use for policy advice at municipalities due to the classification threshold being equal for every image.

The face validity analysis and data validation showed that the utility-based liveability score correlates

most with subjective factors compared to objective factors from the Wijkprofiel of the municipality of Rotterdam. This finding indicates that our model effectively captures the subjective nature of human perception of public spaces. Additionally, substantial correlations between the utility-based liveability score and the amount of green and water in an area highlight the importance of these factors, even though this result is not apparent from the LIME explanations. In the LIME application, no clear type of contribution (solely positive or negative) was found for the greenery superpixels in street view images. But since the problem is present in the superpixel creation, improving the segmentation algorithm could change this observation. The face validity analysis showed the possibility of CVDCMs to analyze cities on preferred liveability and identify areas based on their preferred liveability.

The application of the methodology showed that modelling preferred liveability can provide useful advice for urban policies. Preferred liveability refers to the aspects of an urban environment that individuals prioritize or desire in a given context, reflecting their ideal living conditions. It can provide insights into what people aspire to in their living environments, allowing urban planners to design spaces that better meet residents' desires and improve overall liveability. Policymakers can prioritize interventions that align with residents' aspirations, leading to more effective and sustainable urban development.

# References

- [1] Radhakrishna Achanta et al. "SLIC superpixels compared to state-of-the-art superpixel methods". In: *IEEE transactions on pattern analysis and machine intelligence* 34.11 (2012), pp. 2274–2282.
- [2] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6 (2018), pp. 52138–52160.
- [3] C. Alexander et al. "A pattern language: Towns, buildings, construction". In: (1977). DOI: <https://www-sciencedirect-com.tudelft.idm.oclc.org/science/article/pii/S0169204618310119#b0015>.
- [4] Gulsum Alicioglu and Bo Sun. "A survey of visual analytics for Explainable Artificial Intelligence methods". In: *Computers & Graphics* 102 (2022), pp. 502–520.
- [5] D. Appleyard, M. Gerson, and M. Lintell. *Livable Streets*. University of California Press, 1981.
- [6] A. Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115.
- [7] Hannah M Badland et al. "Can virtual streetscape audits reliably replace physical streetscape audits?" In: *Journal of urban health* 87.6 (2010), pp. 1007–1016.
- [8] S. Barber et al. "Neighborhood Disadvantage and Cumulative Biological Risk Among a Socioeconomically Diverse Sample of African American Adults: An Examination in the Jackson Heart Study". In: *Racial Ethn. Health Dispar* (2016). DOI: <http://dx.doi.org/10.1007/s40615-015-0157-0>.
- [9] Nicholas Berente et al. "Managing artificial intelligence." In: *MIS quarterly* 45.3 (2021).
- [10] Christopher M Bishop. "Pattern recognition and machine learning". In: *Springer google schola* 2 (2006), pp. 1122–1128.
- [11] V. Bruce, P. Green, and M. Georgeson. "Visual perception: Physiology, psychology, & ecology". In: *Psychology Press* (2003).
- [12] John M Bryson. "What to do when stakeholders matter: stakeholder identification and analysis techniques". In: *Public management review* 6.1 (2004), pp. 21–53.
- [13] Ennio Cascetta. "Random utility theory". In: *Transportation Systems Analysis: models and applications* (2009), pp. 89–167.
- [14] T. Cassidy. "Environmental psychology". In: *Psychology Press* (1997). DOI: [https://books.google.nl/books?hl=en&lr=&id=0b3eAQAQBAJ&oi=fnd&pg=PA1&ots=\\_aVGQ6tnGR&sig=hoYqzdsLn3twDA5LZ6VhPnGmdoc&redir\\_esc=y#v=onepage&q&f=false](https://books.google.nl/books?hl=en&lr=&id=0b3eAQAQBAJ&oi=fnd&pg=PA1&ots=_aVGQ6tnGR&sig=hoYqzdsLn3twDA5LZ6VhPnGmdoc&redir_esc=y#v=onepage&q&f=false).
- [15] Roger W Caves. *Encyclopedia of the City*. Routledge, 2004.
- [16] F. Chen S. amd Biljecki. "Automatic assessment of public open spaces using street view imagery". In: *Cities* 137 (2023). DOI: <https://doi.org/10.1016/j.cities.2023.104329>.
- [17] Shuting Chen and Filip Biljecki. "Automatic assessment of public open spaces using street view imagery". In: *Cities* 137 (2023), p. 104329.
- [18] William AV Clark and Valerie Ledwith. "Mobility, housing stress, and neighborhood contexts: evidence from Los Angeles". In: *Environment and Planning A* 38.6 (2006), pp. 1077–1093.
- [19] Dorin Comaniciu and Peter Meer. "Mean shift: A robust approach toward feature space analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.5 (2002), pp. 603–619.
- [20] M. Commers, N. Gottlieb, and G. Kok. "How to change environmental conditions for health". In: *Health Promotion International* (207). DOI: <https://doi-org.tudelft.idm.oclc.org/10.1093/heapro/dal038>.

- [21] European Commission. *AI Act*. 2024. URL: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.
- [22] Marius Cordts et al. "The cityscapes dataset for semantic urban scene understanding". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3213–3223.
- [23] S. van Cranenburgh and F. Garrido-Valenzuela. "Computer vision-enriched discrete choice models, with an application to residential location choice". In: (2023).
- [24] Arun Das and Paul Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey". In: *arXiv preprint arXiv:2006.11371* (2020).
- [25] Lee Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26 (1945).
- [26] Dany Doiron et al. "Predicting walking-to-work using street-level imagery and deep learning in seven Canadian cities". In: *Scientific reports* 12.1 (2022), p. 18380.
- [27] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).
- [28] V. Dover and J. Massengale. *Street design: The secret to great cities and towns*. Wiley, 2013.
- [29] Abhimanyu Dubey et al. "Deep learning the city: Quantifying urban perception at a global scale". In: *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I* 14. Springer. 2016, pp. 196–212.
- [30] William H DuMouchel and Greg J Duncan. "Using sample survey weights in multiple regression analyses of stratified samples". In: *Journal of the American Statistical Association* 78.383 (1983), pp. 535–543.
- [31] Upol Ehsan et al. "Operationalizing human-centered perspectives in explainable AI". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–6.
- [32] G.W. Evans. "The built environment and mental health". In: (2003). DOI: <http://dx.doi.org/10.1093/jurban/jtg063>.
- [33] R. Ewing and S. Handy. "Measuring the Unmeasurable: Urban Design Qualities Related to Walkability". In: *Journal of Urban Design* (2009). DOI: <https://doi.org/10.1080/13574800802451155>.
- [34] Reid Ewing and Susan Handy. "Measuring the unmeasurable: Urban design qualities related to walkability". In: *Journal of Urban design* 14.1 (2009), pp. 65–84.
- [35] Reid H Ewing et al. *Measuring urban design: Metrics for livable places*. Vol. 200. Island Press Washington, DC, 2013.
- [36] Pedro F Felzenszwalb and Daniel P Huttenlocher. "Efficient graph-based image segmentation". In: *International journal of computer vision* 59 (2004), pp. 167–181.
- [37] Prateek Garg et al. "Isdnet: Ai-enabled instance segmentation of aerial scenes for smart cities". In: *ACM Transactions on Internet Technology (TOIT)* 21.3 (2021), pp. 1–18.
- [38] Damien Garreau and Dina Mardaoui. "What does LIME really see in images?" In: *International conference on machine learning*. PMLR. 2021, pp. 3620–3629.
- [39] J. Gehl. "Cities for people". In: (2010).
- [40] J. Gehl and B. Svarre. "How to study public life". In: (2013).
- [41] Zhaoya Gong et al. "Classifying street spaces with street view images for a spatial indicator of urban functions". In: *Sustainability* 11.22 (2019), p. 6424.
- [42] Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009.
- [43] Michael F Goodchild. "Formalizing place in geographic information systems". In: *Communities, neighborhoods, and health: Expanding the boundaries of place*. Springer, 2010, pp. 21–33.
- [44] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

- [45] David Gunning and David Aha. "DARPA's explainable artificial intelligence (XAI) program". In: *AI magazine* 40.2 (2019), pp. 44–58.
- [46] Brookie Guzder-Williams et al. "Intra-urban land use maps for a global sample of cities from Sentinel-2 satellite imagery and computer vision". In: *Computers, Environment and Urban Systems* 100 (2023), p. 101917.
- [47] M. Haan, G.A. Kaplan, and T. Camacho. "Poverty and health. Prospective evidence from the Alameda County Study". In: (1987). DOI: <http://dx.doi.org/10.1093/oxfordjournals.aje.a114637>.
- [48] Pascal Hamm et al. "Explanation matters: An experimental study on explainable AI". In: *Electronic Markets* 33.1 (2023), p. 17.
- [49] Susan L Handy et al. "How the built environment affects physical activity: views from urban planning". In: *American journal of preventive medicine* 23.2 (2002), pp. 64–73.
- [50] C. Harvey. "Measuring Streetscape Design for Livability Using Spatial Data and Methods". In: (2014).
- [51] C. Harvey and L. Aultman-Hall. "Measuring Urban Streetscapes for Livability: A Review of Approaches". In: (2015). URL: <https://www-tandfonline-com.tudelft.idm.oclc.org/doi/full/10.1080/00330124.2015.1065546>.
- [52] C. Harvey et al. "Effects of skeletal streetscape design on perceived safety". In: *Landscape and urban planning* 142 (2015). DOI: 10.1016/j.landurbplan.2015.05.007.
- [53] G. Heath et al. "The effectiveness of urban design and land use and transport policies and practices to increase physical activity: A systematic review". In: *Journal of Physical Activity and Health* (2006).
- [54] Arthur E Hoerl and Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.
- [55] Economist Intelligence. "The Global Liveability Index 2023". In: (2023). URL: <https://www.eiu.com/n/campaigns/global-liveability-index-2023/>.
- [56] Koichi Ito and Filip Biljecki. "Assessing bikeability with street view imagery and computer vision". In: *Transportation research part C: emerging technologies* 132 (2021), p. 103371.
- [57] Paul Jaccard. "Étude comparative de la distribution florale dans une portion des Alpes et des Jura". In: *Bull Soc Vaudoise Sci Nat* 37 (1901), pp. 547–579.
- [58] J. Jacobs. "The death and life of great American cities". In: *New York: Random House* (1961). DOI: [https://www.buurtwijz.nl/sites/default/files/buurtwijz/bestanden/jane\\_jacobs\\_the\\_death\\_and\\_life\\_of\\_great\\_american.pdf](https://www.buurtwijz.nl/sites/default/files/buurtwijz/bestanden/jane_jacobs_the_death_and_life_of_great_american.pdf).
- [59] Abdul Rehman Javed et al. "A survey of explainable artificial intelligence for smart cities". In: *Electronics* 12.4 (2023), p. 1020.
- [60] Ioannis Kakogeorgiou and Konstantinos Karantzalos. "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing". In: *International Journal of Applied Earth Observation and Geoinformation* 103 (2021), p. 102520.
- [61] M. Kashef. "Urban livability across disciplinary and professional boundaries". In: (2016). URL: <https://core.ac.uk/download/pdf/82520225.pdf>.
- [62] George L Kelling and Catherine M Coles. *Fixing broken windows: Restoring order and reducing crime in our communities*. Simon and Schuster, 1997.
- [63] Z. Khorrani et al. "The indicators and methods used for measuring urban liveability: a scoping review". In: *Reviews on Environmental Health* (2021), pp. 397–441. DOI: <https://doi-org.tudelft.idm.oclc.org/10.1515/reveh-2020-0097>.
- [64] G. Kirdar and G. Cagdas. "A decision support model to evaluate liveability in the context of urban vibrancy." In: *International Journal of Architectural Computing* (2022), pp. 528–552. DOI: [doi:10.1177/14780771221121500](https://doi.org/10.1177/14780771221121500).
- [65] Patrick Knab, Sascha Marton, and Christian Bartelt. "DSEG-LIME—Improving Image Explanation by Hierarchical Data-Driven Segmentation". In: *arXiv preprint arXiv:2403.07733* (2024).

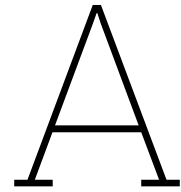
- [66] Stephen Law et al. "Explaining holistic image regressors and classifiers in urban analytics with plausible counterfactuals". In: *International Journal of Geographical Information Science* (2023), pp. 1–22.
- [67] J. Leby and A. Hashim. "Liveability Dimensions and Attributes: Their Relative Importance in the Eyes of Neighbourhood Residents". In: (2010). URL: <https://core.ac.uk/download/pdf/199244965.pdf>.
- [68] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [69] Jiyun Lee, Donghyun Kim, and Jina Park. "A machine learning and computer vision study of the environmental characteristics of streetscapes that affect pedestrian satisfaction". In: *Sustainability* 14.9 (2022), p. 5730.
- [70] K. Leidelmeijer and I. van Kamp. "Kwaliteit van de Leefomgeving en Leefbaarheid". In: *RIVM rapport 630950002/* (2003).
- [71] K. Leidelmeijer and J. Mandemakers. "Leefbaarheid in Nederland 2020". In: (2020).
- [72] B. Li et al. "A bibliometric study and science mapping research of intelligent decision". In: *Cognitive Computation* 14.3 (2022), pp. 989–1008.
- [73] X. Li, C. Ratti, and I. Seiferling. "Mapping urban landscapes along streets using google street view". In: (2017).
- [74] X. Li, C. Zhang, and W. Li. "Does the visibility of greenery increase perceived safety in urban areas? Evidence from the Place Pulse 1.0 Dataset". In: *ISPRS International Journal of Geo-Information* (2015).
- [75] X. Li et al. "Assessing street-level urban greenery using Google Street View and a modified green view index". In: *Urban Forestry & Urban Greening* (2015).
- [76] Zheng Li and Jun Ma. "Discussing street tree planning based on pedestrian volume using machine learning and computer vision". In: *Building and Environment* 219 (2022), p. 109178.
- [77] Xiucheng Liang, Tianhong Zhao, and Filip Biljecki. "Revealing spatio-temporal evolution of urban visual environments with street view imagery". In: *Landscape and Urban Planning* 237 (2023), p. 104802.
- [78] Lin Lin and Anne Vernez Moudon. "Objective versus subjective measures of the built environment, which are most effective in capturing associations with walking?" In: *Health & place* 16.2 (2010), pp. 339–348.
- [79] Zachary C Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.
- [80] Setha M Low and Irwin Altman. "Place Attachment: Human Behavior and Environment. Advances in Theory and Research". In: *Place attachment* (1992), pp. 253–256.
- [81] Arthur Aston Luce, TE Jessop, and George Berkeley. "The Works of George Berkeley, Bishop of Cloyne". In: *British Journal for the Philosophy of Science* 5.17 (1954).
- [82] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).
- [83] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).
- [84] Kevin Lynch. *The image of the city*. MIT press, 1964.
- [85] J. Mandemakers et al. "Leefbaarometer 3.0, Instrumentontwikkeling". In: (2021).
- [86] Charles F Manski. "The structure of random utility models". In: *Theory and decision* 8.3 (1977), p. 229.
- [87] R. Marasinghe et al. "Computer vision applications for urban planning: A systematic review of opportunities and constraints". In: *Sustainable Cities and Society* 100 (2024). DOI: <https://doi.org/10.1016/j.scs.2023.105047>.
- [88] Donella H Meadows. *Thinking in systems: A primer*. chelsea green publishing, 2008.

- [89] Marina Meila. "Comparing Clusterings by the Variation of Information". In: (2003).
- [90] Mehak Maqbool Memon et al. "Unified DeepLabV3+ for semi-dark image semantic segmentation". In: *Sensors* 22.14 (2022), p. 5312.
- [91] Mercer. "Quality of living city ranking". In: (2023). URL: <https://mobilityexchange.mercer.com/Insights/quality-of-living-rankings>.
- [92] Christian Meske et al. "Explainable artificial intelligence: objectives, stakeholders, and future research opportunities". In: *Information Systems Management* 39.1 (2022), pp. 53–63.
- [93] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38.
- [94] Ronald K Mitchell, Bradley R Agle, and Donna J Wood. "Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts". In: *Academy of management review* 22.4 (1997), pp. 853–886.
- [95] C. Montgomery. *Happy city: Transforming our lives through urban design*. Penguin Books, 2013.
- [96] John Montgomery. "Making a city: Urbanity, vitality and urban design". In: *Journal of urban design* 3.1 (1998), pp. 93–116.
- [97] W. Murdoch et al. "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080.
- [98] D. Myers. "Community-relevant measurement of quality of life: A focus on local trends". In: *Urban Affairs Quarterly* (1987). DOI: <https://journals.sagepub.com/doi/pdf/10.1177/004208168702300107>.
- [99] Nikhil Nikhil Deepak Naik. "Visual urban sensing: understanding cities through computer vision". PhD thesis. Massachusetts Institute of Technology, 2017.
- [100] Ladan Najafizadeh and Jon E Froehlich. "A feasibility study of using Google street view and computer vision to track the evolution of urban accessibility". In: *Proceedings of the 20th international ACM SIGACCESS conference on computers and accessibility*. 2018, pp. 340–342.
- [101] J Nasar. "The evaluative image of the city". In: *Journal of the American Planning Association* 56.1 (1990), pp. 41–53.
- [102] J. Nasar. "Perception, cognition, and evaluation of urban places". In: *Public places and spaces* (1989). DOI: [https://doi.org/10.1007/978-1-4684-5601-1\\_3](https://doi.org/10.1007/978-1-4684-5601-1_3).
- [103] Jack L Nasar. "The evaluative image of the city". In: *Journal of the American Planning Association* 56.1 (1990), pp. 41–53.
- [104] *OECD Better Life Index*. 2023. URL: [https://www.oecdbetterlifeindex.org/#/111111111111111](https://www.oecdbetterlifeindex.org/#/1111111111111).
- [105] Vicente Ordonez and Tamara L Berg. "Learning high-level judgments of urban perception". In: *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*. Springer. 2014, pp. 494–510.
- [106] P. Owens. "Neighborhood form and pedestrian life: Taking a closer look". In: *Landscape and Urban Planning* (1993). DOI: [https://doi.org/10.1016/0169-2046\(93\)90011-2](https://doi.org/10.1016/0169-2046(93)90011-2).
- [107] Karl Pearson. "VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia". In: *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 187 (1896), pp. 253–318.
- [108] Lorenzo Porzi et al. "Predicting and understanding urban perception with convolutional neural networks". In: *Proceedings of the 23rd ACM international conference on Multimedia*. 2015, pp. 139–148.
- [109] W. Qiu et al. "Subjectively Measured Streetscape Perceptions to Inform Urban Design Strategies for Shanghai". In: *ISPRS International Journal of Geo-Information*. 10 (2021). DOI: <https://doi.org/10.3390/ijgi10080493>.
- [110] Muhammad Rashid et al. "Using Stratified Sampling to Improve LIME Image Explanations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 13. 2024, pp. 14785–14792.

- [111] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ““ Why should I trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.
- [112] Rijksoverheid. “Inkomsten gemeenten en provincies”. In: (2024). URL: <https://www.rijksoverheid.nl/onderwerpen/financien-gemeenten-en-provincies/inkomsten-gemeenten-en-provincies#:~:text=Gemeenten%20en%20provincies%20hebben%20eigen,provincies%20geld%20van%20de%20rijksoverheid..>
- [113] Rijksoverheid. “Taken van een gemeente”. In: (2024). URL: <https://www.rijksoverheid.nl/onderwerpen/gemeenten/taken-gemeente>.
- [114] Horst WJ Rittel and Melvin M Webber. “Dilemmas in a general theory of planning”. In: *Policy sciences* 4.2 (1973), pp. 155–169.
- [115] Gemeente Rotterdam. “Wijkprofiel Rotterdam 2024”. In: (2024). URL: <https://wijkprofiel.rotterdam.nl/nl/2024/rotterdam>.
- [116] B. Saelens, J. Sallis, and L. Frank. “Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures”. In: *Annals of Behavioral Medicine*, 25 (2003). DOI: [https://doi.org/10.1207/S15324796ABM2502\\_03](https://doi.org/10.1207/S15324796ABM2502_03).
- [117] Paul A Samuelson. “Consumption theory in terms of revealed preference”. In: *Economica* 15.60 (1948), pp. 243–253.
- [118] Ludwig Schallner et al. “Effect of superpixel aggregation on explanations in lime—a case study with biological data”. In: *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*. Springer. 2020, pp. 147–158.
- [119] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.
- [120] C. Shafer, B. Lee, and S. Turner. “A tale of three greenway trails: User perceptions related to quality of life”. In: *Landscape and Urban Planning* (2000). DOI: [https://staff.washington.edu/kwolf/Archive/Classes/ESRM304\\_SocSci/304%20Soc%20Sci%20Lab%20Articles/Shofer\\_2000.pdf](https://staff.washington.edu/kwolf/Archive/Classes/ESRM304_SocSci/304%20Soc%20Sci%20Lab%20Articles/Shofer_2000.pdf).
- [121] Wenzhong Shi. “Introduction to urban sensing”. In: *Urban Informatics* (2021), pp. 311–314.
- [122] Wesley G Skogan. *Disorder and decline: Crime and the spiral of decay in American neighborhoods*. Univ of California Press, 1992.
- [123] Tim Heinrich Son et al. “Algorithmic urban planning for smart and sustainable development: Systematic review of the literature”. In: *Sustainable Cities and Society* (2023), p. 104562.
- [124] Douglas Steinley. “Properties of the hubert-arable adjusted rand index.” In: *Psychological methods* 9.3 (2004), p. 386.
- [125] Emily Talen and Julia Koschinsky. “Compact, walkable, diverse neighborhoods: Assessing effects on residents”. In: *Housing policy debate* 24.4 (2014), pp. 717–750.
- [126] J. Tang and Y. Long. “Measuring visual quality of street space and its temporal variation: Methodology and its application in the Hutong area in Beijing”. In: *Landscape and Urban Plannin* (2019). DOI: <https://doi.org/10.1016/j.landurbplan.2018.09.015>.
- [127] S. Thompson and J. Kent. “Healthy Built Environments Supporting Everyday Occupations: Current Thinking in Urban Planning”. In: (2014). DOI: [dx.doi.org/10.1093/oxfordjournals.aje.a114637](https://doi.org/10.1093/oxfordjournals.aje.a114637).
- [128] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2003.
- [129] Yi-Fu Tuan. *Landscapes of fear*. U of Minnesota Press, 2013.
- [130] Yi-Fu Tuan. *Space and place: The perspective of experience*. U of Minnesota Press, 1977.
- [131] V. ujjatha, G. Lavanya, and R. Prakash. “Quantifying Liveability Using Survey Analysis and Machine Learning Model”. In: *Landscape and Urban Plannin* (2023). DOI: <https://doi.org/10.3390/su15021633>.



- [132] Bas HM Van der Velden et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". In: *Medical Image Analysis* 79 (2022), p. 102470.
- [133] Ann Van Herzele and Torsten Wiedemann. "A monitoring tool for the provision of accessible and attractive urban green spaces". In: *Landscape and urban planning* 63.2 (2003), pp. 109–126.
- [134] Andrea Vedaldi and Stefano Soatto. "Quick shift and kernel methods for mode seeking". In: *Computer Vision—ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12–18, 2008, Proceedings, Part IV 10*. Springer. 2008, pp. 705–718.
- [135] Tom Vermeire et al. "Explainable image classification with evidence counterfactual". In: *Pattern Analysis and Applications* 25.2 (2022), pp. 315–335.
- [136] Luc Vincent and Pierre Soille. "Watersheds in digital spaces: an efficient algorithm based on immersion simulations". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 13.06 (1991), pp. 583–598.
- [137] Kentaro Wada. *Labelme: Image Polygonal Annotation with Python*. DOI: 10.5281/zenodo.5711226. URL: <https://github.com/wkentaro/labelme>.
- [138] Ralf Weber, Jörg Schnier, and Thomas Jacobsen. "Aesthetics of streetscapes: Influence of fundamental properties on aesthetic judgments of urban space". In: *Perceptual and motor skills* 106.1 (2008), pp. 128–146.
- [139] Victor Wiley and Thomas Lucas. "Computer vision and image processing: a paper review". In: *International Journal of Artificial Intelligence Research* 2.1 (2018), pp. 29–36.
- [140] Bernd W Wirtz, Jan C Weyerer, and Carolin Geyer. "Artificial intelligence and the public sector—applications and challenges". In: *International Journal of Public Administration* 42.7 (2019), pp. 596–615.
- [141] X. Xu et al. "Associations between Street-View Perceptions and Housing Prices: Subjective vs. Objective Measures Using Computer Vision and Machine Learning Techniques". In: *Remote Sensing* 14 (2022). DOI: <https://doi.org/10.3390/rs14040891>.
- [142] Yao Yao et al. "Discovering the homogeneous geographic domain of human perceptions from street view images". In: *Landscape and Urban Planning* 212 (2021), p. 104125.
- [143] Tan Yigitcanlar et al. "Artificial intelligence in local government services: Public perceptions from Australia and Hong Kong". In: *Government Information Quarterly* (2023), p. 101833.
- [144] Li Yin and Zhenxin Wang. "Measuring visual enclosure for street walkability: Using machine learning algorithms and Google Street View imagery". In: *Applied geography* 76 (2016), pp. 147–153.
- [145] Fan Zhang et al. "Measuring human perceptions of a large-scale urban region using machine learning". In: *Landscape and Urban Planning* 180 (2018), pp. 148–160.



## Globally known liveability frameworks

Kashef [61] reviews three of the most well-known frameworks for liveability, which are used to determine liveability on a global nationwide scale. He reviews 3 globally used liveability measurements: the EIU liveability ranking, Mercer quality of living survey, and the OECD BLI.

The EIU liveability ranking [55] stands out as the most comprehensive and extensive among all systems assessing quality of life. With 40 liveability indicators grouped into five weighted categories, namely: stability (25%), healthcare (20%), culture and environment (25%), education (10%), and infrastructure (20%). Cities receive ratings ranging from 0 (worst) to 100 (ideal) based on their performance in these dimensions. The stability measure encompasses factors like crime rates and threats of civil unrest. In the healthcare category, cities score higher by offering quality, affordable private/public medical services. Culture and environment incorporate indicators related to things such as climate, freedom, and air quality. Education indicators evaluate the provision and quality of private/public educational institutions across various learning levels. The infrastructure category includes diverse indicators assessing the quality of local road networks, connectivity, and the availability of water, energy, and quality housing for example. EIU employs various data collection and measurement tools, encompassing raw quantitative data, public opinion surveys, and interviews with a broad spectrum of professionals. Surveys and interviews are conducted worldwide, with representative samples of respondents drawn from Asia (30%), the Americas (30%), Europe (30%), and other parts of the world (10%).

The Mercer Quality of Living Survey [91] measures the liveability in 460 cities worldwide. Quality of living is evaluated through 39 indicators organized into 10 classifications. They are the sociopolitical environment (crime, safety, and stability), economics (banking regulations and services), socio-cultural environment (media, censorship, and personal freedom), health (private and public services, air quality, sanitation, and waste disposal), education (private and public), utilities (transportation, traffic, and services), recreational facilities (restaurants, theaters, sports, and leisure), market (availability of goods), housing, and natural environment (climate, natural calamities, and weather extremes). It allows also for comparison between cities, allowing the user to put more emphasis on specific indicators to compare them and give them more weight.

The OECD Better Life Index [104] measures and compares the quality of life among 34 OECD member countries. The OECD launched this in 2011 to synchronize well-being and quality-of-life measures with the recommendations of the Commission on the Measurement of Economic Performance and Social Progress (the Stiglitz-Sen-Fitoussi Commission). This index measures liveability with indicators such as housing, income, employment, social support systems, education, environment, health, governance, life satisfaction, safety, and work-life balance. The data used comes from national records and public surveys.

As can be seen, these qualities of life measurements in cities or nations use a vast amount of indicators for estimating liveability, ranging from social to political factors. Therefore, it is dependent on a vast amount of data, with all factors being complex. This results in these frameworks not easily applying to other research. Since in this research, the focus is on street view images, and no other input data, these frameworks are too complex and large for this research to use. Therefore, liveability research on the neighborhood level is needed.

## A.1. Leefbarometer

A well-known Dutch liveability measurement on the city level is the Leefbaarometer project Leidelmeijer and Mandemakers [71]. The Leefbaarometer surveys the liveability of neighborhoods across the Netherlands, where it uses its factors for determining liveability. The project aims to model liveability using variables applicable nationwide, such as housing quality and greenspace proximity. However, the use of manually collected data, such as surveys, poses challenges in upscaling and replicating results. Despite these challenges, linking survey data to empirical and statistical information holds the potential to enhance our understanding of the factors contributing to the liveability of cities on a broader scale. It is the most well-known Dutch liveability measurement tool and is used nationwide for information and policy interventions.

The Leefbaarometer is an instrument that provides an estimation of liveability at a low scale level (100x100 meters) for the entire Netherlands based on various characteristics of the living environment. The purpose of the Leefbaarometer is to offer an overview of where liveability is expected to be either good or poor and how it is evolving (thus timestamped liveability). The instrument serves as both a signaling and monitoring tool, offering a statistical prediction of local liveability. Their definition of liveability is taken from a study by Leidelmeijer and Kamp [70]: 'Liveability is the degree to which the environment aligns with the requirements and desires expressed by humans.' The Leefbaarometer utilizes two models: the residents' judgment model and the housing market behavior model. The residents' judgment model is based on residents' opinions about how pleasant it is to live in their neighborhood, derived from the Housing Survey Netherlands (WoON) with 67 thousand respondents in 2018. The behavior model analyzes the appreciation for an environment based on over 700 thousand transactions in the housing market from 2017 to 2019. The Leefbaarometer maps the weighted sum of various environmental features, grouped into five dimensions. The factors that influence liveability in the Netherlands the most are nuisance and safety, as well as facilities, which carry the most weight in the model of the Leefbaarometer, followed by housing stock, social cohesion, and the physical environment. In our research, we use street view images for liveability measurement, and therefore it is important to emphasize that in these images, safety and facilities are hard to be included. These images are more closely related to the physical environment, which is deemed the least important of the 5 features in the Leefbaarometer.

It is important to emphasize that the Leefbaarometer does not provide recommendations on how liveability can be improved. This is because not all relevant environmental characteristics are included in the model, primarily due to a lack of comparable data for both local and national levels. The Leefbaarometer indicates liveability per residential area in a specific year, normalized to the Housing Survey Netherlands (WoON) 2018 or as a deviation from the average in the Netherlands in 2018.

In general, liveability is defined as the degree to which the environment aligns with the requirements and desires expressed by humans. Various definitions emphasize aspects such as the quality of place, social and physical factors improving the standard of living, and the relationship between people and their environment. The Leefbaarometer adheres to the definition that liveability is the degree to which the environment aligns with the requirements and desires of people. In their literature review on liveability done before the modeling, the following environmental characteristics related to neighborhood liveability are identified: health (physical and mental well-being of residents), aesthetics/attractiveness of the neighborhood, well-being (time spent, meaningfulness, etc. for residents), positive social relationships (feeling at home, cohesion, residents' trust), opportunities (employment, education, healthcare, influence in the neighborhood), and (perceived) safety of the neighborhood. Based on the literature, definitions of liveability, and these environmental factors, they establish five dimensions influencing neighborhood liveability: the physical environment, housing stock, facilities, social cohesion, and nui-

sance and insecurity. These five dimensions are also used in the models for liveability measurements. In the weighting of these models, facilities, nuisance, and insecurity carry the most weight, followed by housing stock, then social cohesion, and finally the physical environment. The physical environment encompasses the natural environment, infrastructure, businesses, and public spaces. The natural environment includes factors such as environmental quality (air, water, soil, noise), climate, natural disasters, and landscapes. Environmental quality directly impacts our health (Commers, Gottlieb, and Kok [20]). Climate, and climate change, influence liveability in terms of temperature and precipitation. Natural disasters include the possibility of a nearby dike breaking, for example. Landscapes relate to the presence of greenery (trees and forests) in the neighborhood. Infrastructure includes things like roads, train tracks, and tram lines in the vicinity. Housing stock concerns the homes in the neighborhood; attractive or unattractive homes can make a neighborhood desirable or undesirable, and affordability is also a factor. Facilities refer to the minimum distance to a facility, as well as the size and diversity of offerings in the nearby area. Facilities include things like public transportation, shops, hospitals, education, etc. Social cohesion involves the residents in a neighborhood, their relationships, and the opportunities available in the neighborhood. Nuisance and insecurity cover natural disasters, as in the physical environment, but also address traffic safety and crime, for example (Mandemakers et al. [85]).

# B

## Actor Analysis Public Space Perception Municipality of Rotterdam

In this actor analysis, we investigate the various stakeholders and actors involved in the experience of public spaces within Rotterdam. The analysis focuses on identifying which municipal bodies are engaged in or have an impact on the experience of public spaces, both directly and indirectly.

### B.1. Cluster Stadsbeheer

Stadsbeheer is responsible for managing Rotterdam, dedicating efforts to all aspects of the public space. They continuously work on maintaining and improving public spaces, directly impacting how citizens experience them. Within Stadsbeheer, various departments are actively and indirectly focused on enhancing the public space experience in Rotterdam. Consequently, Stadsbeheer effectively manages all elements in the city's public space, which collectively and individually influence perception.

Currently, Stadsbeheer evaluates every 'asset' in the city based on four criteria: condition, cleanliness, integrity, and safety. These assessments are made based on internal insights and CROW guidelines (more on this later). Previously, municipal management was primarily risk-based, focusing on the condition of individual objects. However, the approach now strives for value-driven management, emphasizing added value relative to other aspects, such as liveability. This policy considers values like sustainability and circularity.

In urban development projects, Stadsontwikkeling (SO) and Stadsbeheer (SB) often collaborate intensively. Public space management is an ongoing process, making Stadsbeheer potentially more focused on perception than Stadsontwikkeling. Stadsbeheer also actively engages in dialogue with neighbourhood residents.

#### B.1.1. Afdeling Schone en Circulaire Stad

Within Stadsbeheer, the Schone en Circulaire Stad department ensures the city is free of litter. Clean streets contribute to a sense of safety and aesthetic appeal, enhancing the public space experience. This department is committed to maintaining a tidy appearance of the streets, removing graffiti, collecting waste, organizing bulky waste collection services, and removing objects that detract from the street's beauty. A clean environment not only promotes a sense of safety but also encourages social interaction and better use of outdoor spaces.

#### B.1.2. Afdeling Toezicht, Handhaving, Parkeren en Markten

Stadsbeheer is also responsible for city surveillance and enforcement, contributing to the safety and aesthetics of Rotterdam. Public lighting management falls under their responsibility, directly influencing how citizens perceive the safety of public spaces. The municipal surveillance department focuses on maintaining public order. They have a 'flying squad' to tackle illegal dumping near waste containers and

monitor CCTV in public spaces to deploy enforcers when necessary. This department has a control room in contact with all enforcers in the city to respond efficiently to situations. By maintaining public order and penalizing violations, they indirectly contribute to the city's aesthetics, ultimately improving public space perception. This department also includes Traffic Control, responsible for monitoring traffic in Rotterdam, contributing to the city's mobility and public space perception. Supervisors are active at markets, ensuring compliance with rules and regulations. The Parking department handles parking policy implementation, both policy aspects and practical execution fall under Stadsbeheer. They manage parking garages, and bike storage, and enforce parking rules in public spaces. Addressing problematic parking and enforcing parking rules also impact public space perception. The enforcers of this department are active in the city to maintain public order and prevent outdoor space deterioration, directly influencing its experience. They act as hosts, assisting and engaging with people to maintain neighbourhood liveability. Enforcers have the authority to issue fines, similar to the police, though the police are not part of the Rotterdam municipality.

### **B.1.3. Afdeling Openbare Werken**

Openbare Werken focuses on various aspects of public space in Rotterdam.

### **B.1.4. Afdeling Openbare Werken: Stedelijk Beheer**

Stedelijk Beheer is concerned with creating a vision for how to manage. So this is how they determine the strategy to move towards value-based management for Stadsbeheer, for example. They set how the public space is managed, and for example department of Gebiedsbeheer operates on this vision.

### **B.1.5. Afdeling Openbare Werken: Stedelijk Beheer: AMOR**

The Asset Management Openbare Ruimte (AMOR) department manages all assets in Rotterdam, ensuring they function well and look well-maintained. This contributes to the functionality and aesthetics of public spaces. They do this with special attention to 3 things concerning the public space: green management, playing and the CROW. Green management in Rotterdam includes parks, trees, and green spaces. This encompasses the maintenance and development of green areas, contributing to biodiversity and city well-being. Green spaces enhance the public space experience and increase the sense of safety. Parks play a crucial role in the health, greening, and social integration of the city, influencing the overall public space experience. Green policy is primarily shaped by Stadsontwikkeling (SO) and Stadsbeheer (SB), with input from Maatschappelijke Ontwikkeling (MO). The green agenda, "Rotterdam Gaat Voor Groen," includes policies, action plans, and goals, including climate adaptation measures to handle extreme weather conditions. AMOR is also responsible for managing playgrounds in the city. These playgrounds serve as meeting points for children and residents, promoting social cohesion and contributing to a safer neighbourhood.

### **B.1.6. Afdeling Openbare Werken: Stedelijk Beheer: CROW**

The CROW system is used as a measurement method to assess the objective quality of public spaces. This method is applied to evaluate the condition of all public facilities, informing decisions on renovations or improvements. The measurements also serve as policy guidelines, aiming to meet certain minimum quality standards to ensure the functionality and aesthetics of public spaces.

### **B.1.7. Afdeling Openbare Werken: Gebiedsbeheer**

Gebiedsbeheer acts as a link within Rotterdam's areas. Based in various district offices, they serve as the direct representatives of the neighbourhood. This includes area and district managers, who play a key role in facilitating participation and providing a contact point for residents with the municipality. They highlight community needs and concerns to relevant parties.

## **B.2. Cluster Maatschappelijke ontwikkeling**

The Maatschappelijke Ontwikkeling cluster acts as a policy body for the municipality, setting policies for public spaces in Rotterdam. Maatschappelijke Ontwikkeling focuses on various policy areas, including sports, culture, education, public health, community support, youth policy, homelessness care, welfare, and care. While they develop policies, implementation often lies elsewhere, with external parties or other municipal departments.

For example, they set policies regarding social influences in neighbourhoods, such as establishing neighbourhood councils and facilitating citizen initiatives. They also create policies on poverty reduction and health in neighbourhoods, determining available facilities and promoting community health, such as by facilitating sports opportunities in public spaces. Maatschappelijke Ontwikkeling works with Stadsbeheer and Stadsontwikkeling to design public spaces that promote a healthy lifestyle. This includes policies on sports facilities and encouraging healthy living, benefiting both health and social cohesion. They also develop cultural policies, enhancing social interaction and community spirit within neighbourhoods.

Maatschappelijke Ontwikkeling is thus indirectly closely involved in the public space experience, contributing not only to its design but also by facilitating health and social interactions in the neighbourhood.

## **B.3. Cluster Diensverlening**

The Dienstverlening cluster manages all citizen contact with the municipality. Therefore, Dienstverlening is primarily socially involved in the public space experience in Rotterdam. For example, the municipality's customer service falls under Dienstverlening. Dienstverlening thus brings together many aspects and people in the neighbourhood, based on policies created by Maatschappelijke Ontwikkeling.

### **B.3.1. MELDR App**

The MELDR app is a service platform that enables citizens to report issues in public spaces via a convenient app. This includes problems like litter, damaged objects, or other irregularities. By allowing citizens to make direct reports, they contribute to improving public spaces.

### **B.3.2. Wijken, Participatie, Stadsarchief**

The Dienstverlening department plays a significant role in executing public space tasks, in collaboration with citizens.

### **B.3.3. Wijken, Participatie, Stadsarchie: Wijkregisseurs**

District managers have direct contact with neighbourhoods, acting as a link between municipal services, the Rayon Director (a type of area manager in Rotterdam), and the neighbourhood networker. They work closely with neighbourhood initiatives and operate from the Wijkhub, a location in the neighbourhood accessible to citizens for activities that benefit liveability.

### **B.3.4. Wijken, Participatie, Stadsarchie: Wijkhubs**

The Wijkhubs provide a direct location in the neighbourhood where municipal employees and citizens can meet and collaborate on neighbourhood-oriented activities.

### **Wijken, Participatie, Stadsarchie: CityLab**

CityLab is a platform where residents can submit initiatives and apply for subsidies to support them. This platform promotes citizen participation and supports various initiatives, from cleanup actions to creating green spaces. It contributes to a vibrant and engaged community.

### **B.3.5. Wijken, Participatie, Stadsarchie: Wijk Aan Zet**

Wijk Aan Zet is an initiative directly connected to the neighbourhood council and other Dienstverlening and Stadsbeheer departments. It facilitates neighbourhood initiatives and integrates them into municipal policy, aiming to increase municipal involvement in local initiatives and improve neighbourhood liveability. It acts as a bridge between the municipality and the citizens.

## **B.4. Cluster Stadsontwikkeling**

The Department of Urban Development (Stadsontwikkeling) in Rotterdam plays a pivotal role in shaping the city's growth and development. Their responsibilities encompass several critical areas that directly impact the urban landscape and public space experience. Stadsontwikkeling oversees urban planning initiatives, determining where and how development occurs within Rotterdam. This includes zoning

regulations, land use planning, and the integration of new housing projects. By strategically planning urban areas, they aim to optimize living conditions, promote sustainable growth, and enhance the overall quality of life for residents. The department also focuses on mobility solutions, which are crucial for ensuring efficient transportation networks within the city. This involves planning and implementing infrastructure projects such as roads, bike lanes, and public transportation systems. By improving mobility, Stadsontwikkeling enhances accessibility to public spaces, making them more functional and accommodating for all citizens. Economic vitality is another key area of focus. Stadsontwikkeling promotes economic growth by supporting businesses and commercial activities within public spaces. This includes fostering an environment where shops, restaurants, and other amenities thrive, thereby creating job opportunities and contributing to local prosperity. Aesthetic considerations are integral to their work, as they are responsible for the visual and spatial design of public areas. This includes parks, squares, streetscapes, and other communal spaces. By implementing cohesive design principles and ensuring high-quality standards, they aim to create attractive environments that enhance the overall experience and perception of public spaces. Managing municipal real estate assets is also under their purview. This involves overseeing properties owned by the city, which significantly influences the aesthetic appeal and functional use of the built environment. By maintaining and developing these properties strategically, Stadsontwikkeling contributes to the overall attractiveness and liveability of Rotterdam.

#### **B.4.1. Afdeling Stedelijk Inrichting**

The Stedelijk Inrichting (Urban Design) Department guides the spatial design of the city of Rotterdam. They contribute to a safe, physical, and inclusive living environment, ensuring an attractive and clean outdoor space. Their focus is on how public spaces are organized and maintained to improve the city's overall liveability.

#### **B.4.2. Afdeling Gebiedsontwikkeling**

The Gebiedsontwikkeling (Area Development) Department is responsible for the development of various regions within Rotterdam. Their goal is to create an attractive living environment and ensure an adequate housing supply in the city. They also focus on the planning and utilization of public spaces to enhance the quality of life for residents.

#### **B.4.3. Afdeling Economie en Duurzaamheid**

The Economy and Sustainability Department within Stadsontwikkeling manages the business aspects of the city, including the entrepreneurs' desk, retail policies, and hospitality permits. They also work on enhancing the city's sustainability. By fostering a vibrant economic environment, they improve the liveliness and attractiveness of public spaces. Shops and restaurants contribute to neighbourhood amenities, positively affecting public space perception.

### **B.5. Afdeling Onderzoek en Business Intelligence (OBI)**

The Department of Research and Business Intelligence (OBI) at the Municipality of Rotterdam plays a crucial role in providing knowledge and information essential for shaping and adjusting policies. They fall under the Innovation, Information, Facilities and Research cluster. Their work directly influences how residents perceive and understand public spaces. OBI conducts various forms of research, ranging from in-depth studies to gathering current data on key topics such as economy, youth, health, income, housing, traffic, education, and safety. This research can then be used by other clusters for policies. They analyze data thoroughly and make it accessible through figures, datasets, and dashboards. This enables policymakers, researchers, and the public to understand trends and make decisions based on factual information. Additionally, OBI develops knowledge networks to facilitate the sharing of expertise and best practices within and beyond the municipal context. They also provide advice on effective strategies for utilizing data, crucial for optimizing policy measures and achieving societal goals. They provide both objective (factual data) and subjective (opinions and perceptions) information, as seen in the Rotterdam District Profile. This helps in understanding the liveability, safety, and social cohesion in different neighbourhoods, contributing to a nuanced understanding of how residents experience their immediate surroundings. Moreover, OBI offers insights into health data and lifestyle trends, influencing the quality of life and well-being of citizens. By sharing this information, policymakers can implement



measures that promote health and well-being in public spaces.

### **B.5.1. Onderzoeksteam**

This is a team from OBI that researches enquiries from the municipality.

### **B.5.2. Advanced Analytics**

Advanced analytics is the department where there is the greatest technical knowledge about data and models. Through data engineering, they create surveys, from tabular data to image data. They do the more technical research like deeper data analysis and model building. This department also contains the technical knowledge needed to use computer vision models. They mostly work on a project basis.

## **B.6. Wijkraad**

The neighbourhood council consists of residents who discuss initiatives and issues within the neighbourhood. Acting as a liaison between citizens and the municipality, they address local concerns and advocate for improvements in public spaces. This social impact indirectly enhances the perception and use of these areas.

## **B.7. Gemeenteraad, het college burgemeester en wethouders**

The municipal council, along with the mayor and aldermen, represents all Rotterdam residents and forms the city's highest administrative body. They set policies that the municipality must follow, greatly influencing public space management and development. Their decisions shape the long-term vision and implementation of projects affecting public spaces.

## **B.8. Dutch National Government**

The national government influences municipalities through budget allocations and regulations. Financial support and legislative frameworks provided by the national government impact local policies and projects, indirectly affecting the public space experience. For instance, funding for infrastructure projects can enhance accessibility and safety in public spaces.

## **B.9. European Commission**

The European Commission influences municipalities through various directives and funding programs. This includes environmental regulations, urban development funds, and policies like the GDPR and AI Act. These regulations ensure data protection and ethical AI use in public space modelling, impacting how human perception is studied and improved.

## **B.10. Citizens of Rotterdam**

Residents are both influencers and beneficiaries of public space developments. They shape public space perception through their use and feedback. Citizen participation in municipal decisions, facilitated by tools like the MELDR app, ensures that public spaces meet community needs. Their involvement in neighbourhood councils and initiatives like CityLab fosters a sense of ownership and community, enhancing the overall experience of public spaces. Citizens influence the municipality and national government through participation in local governance, feedback mechanisms, and public consultations. Their needs and preferences guide policy-making and project implementation, ensuring that public spaces are designed and maintained in ways that reflect community values and enhance liveability.

C

Utility based liveability analysis of all  
selected areas

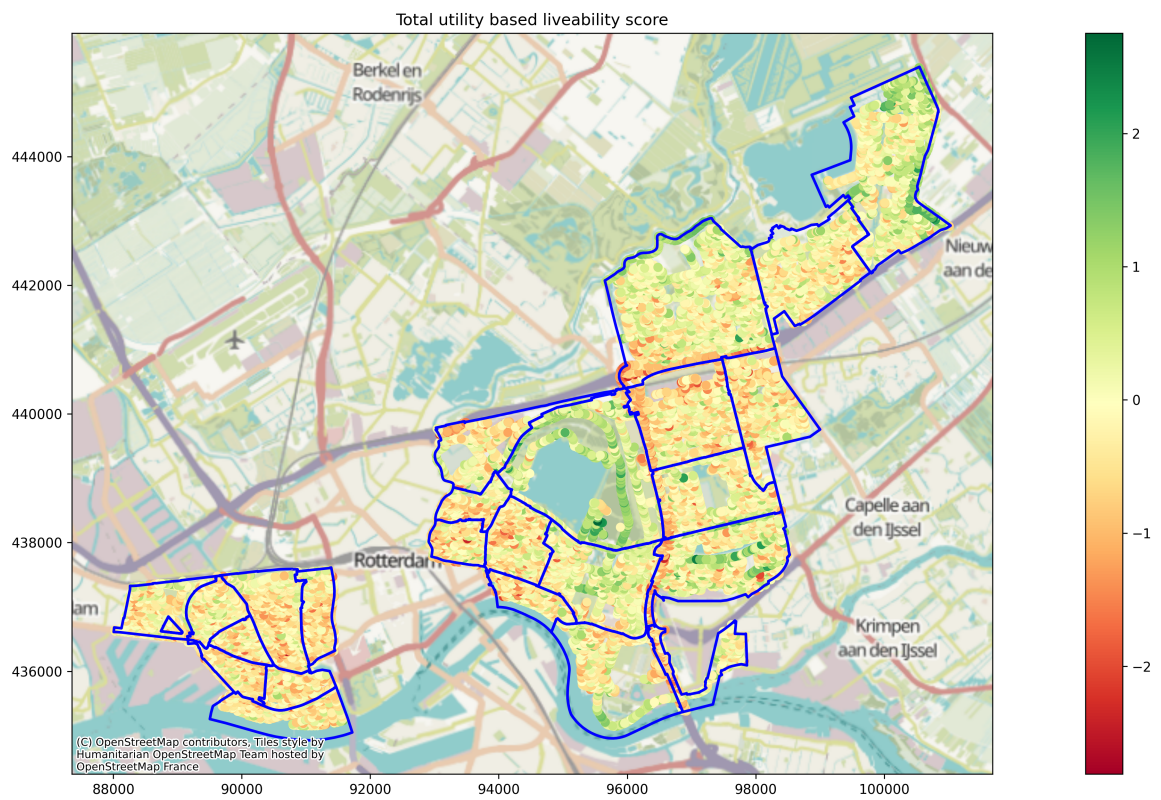
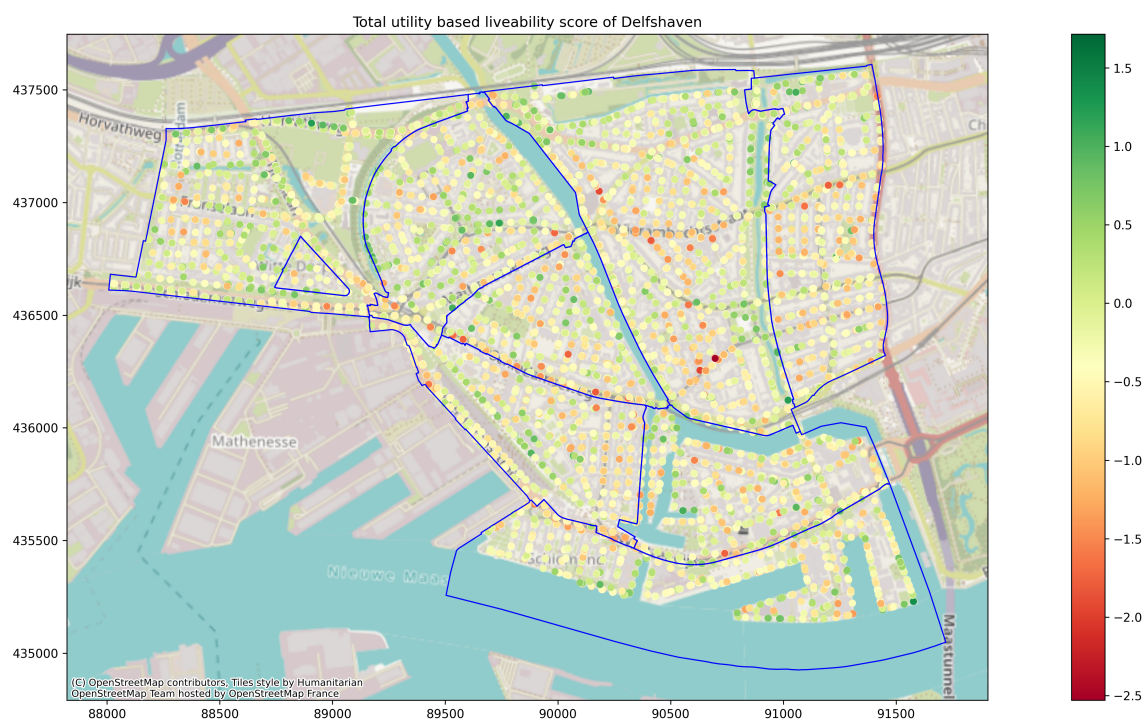


Figure C.1: Utility-based liveability score mapped

D

Additional neighbourhood  
utility-based liveability plots



**Figure D.1:** Delfshaven utility-based liveability score per point

### Total utility based liveability score of Prins-Alexander

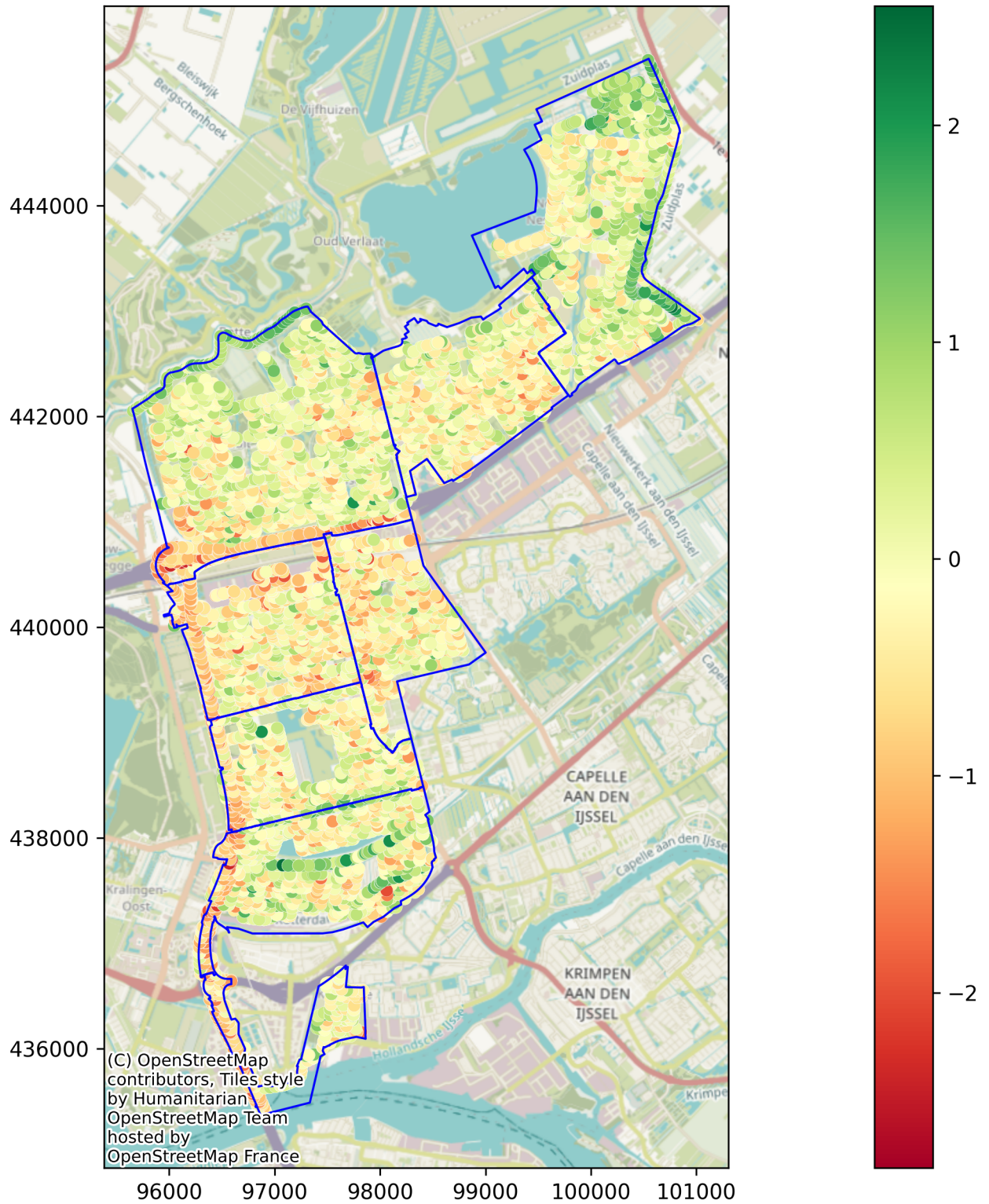


Figure D.2: Prins-Alexander utility-based liveability score per point

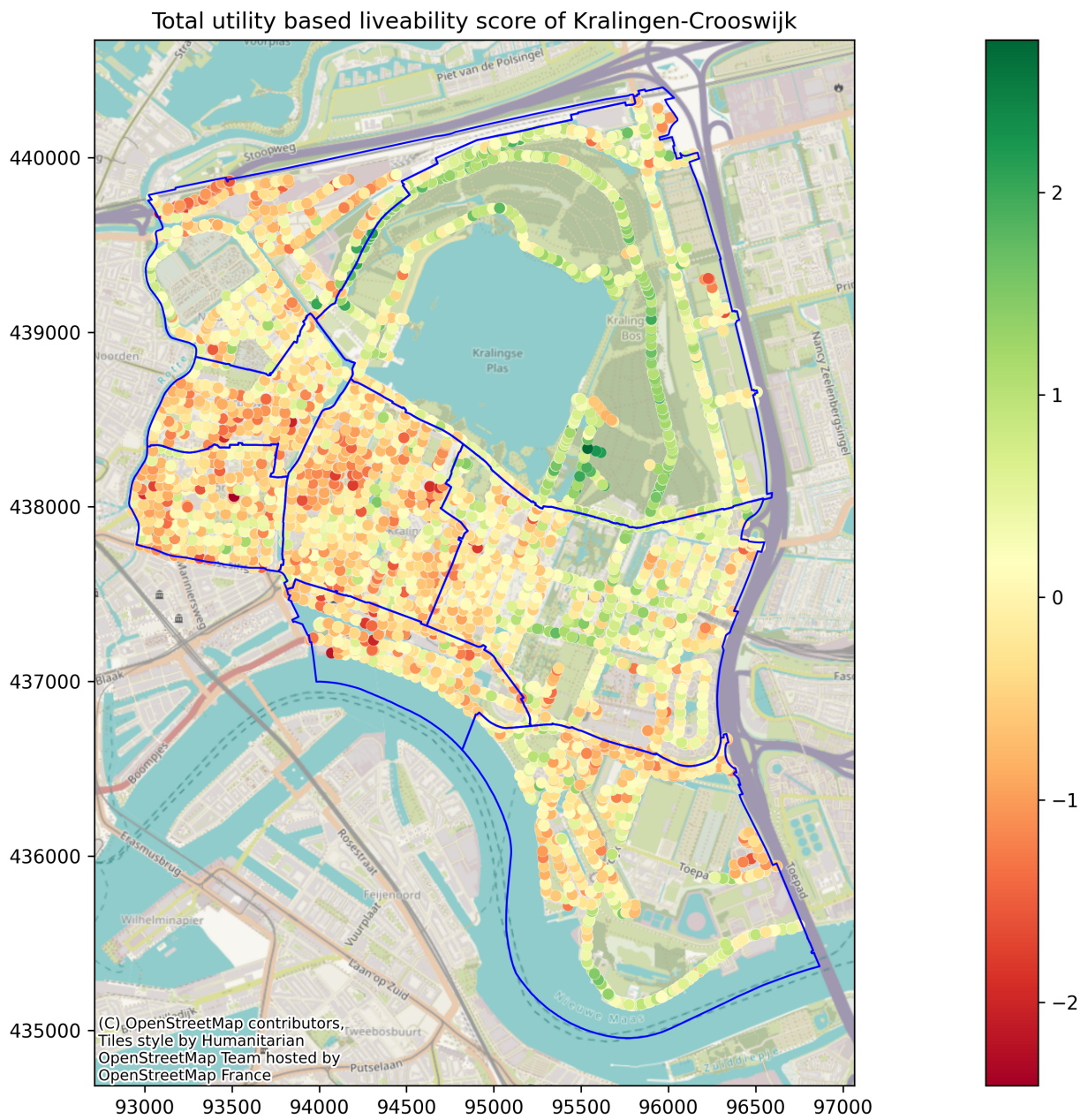


Figure D.3: Kralingen Crooswijk utility-based liveability score per point

E

Prins Alexander en Kralingen  
Crooswijk mapped together



Total utility based liveability score of Kralingen Crooswijk and Prins Alexander per zipcode level 5

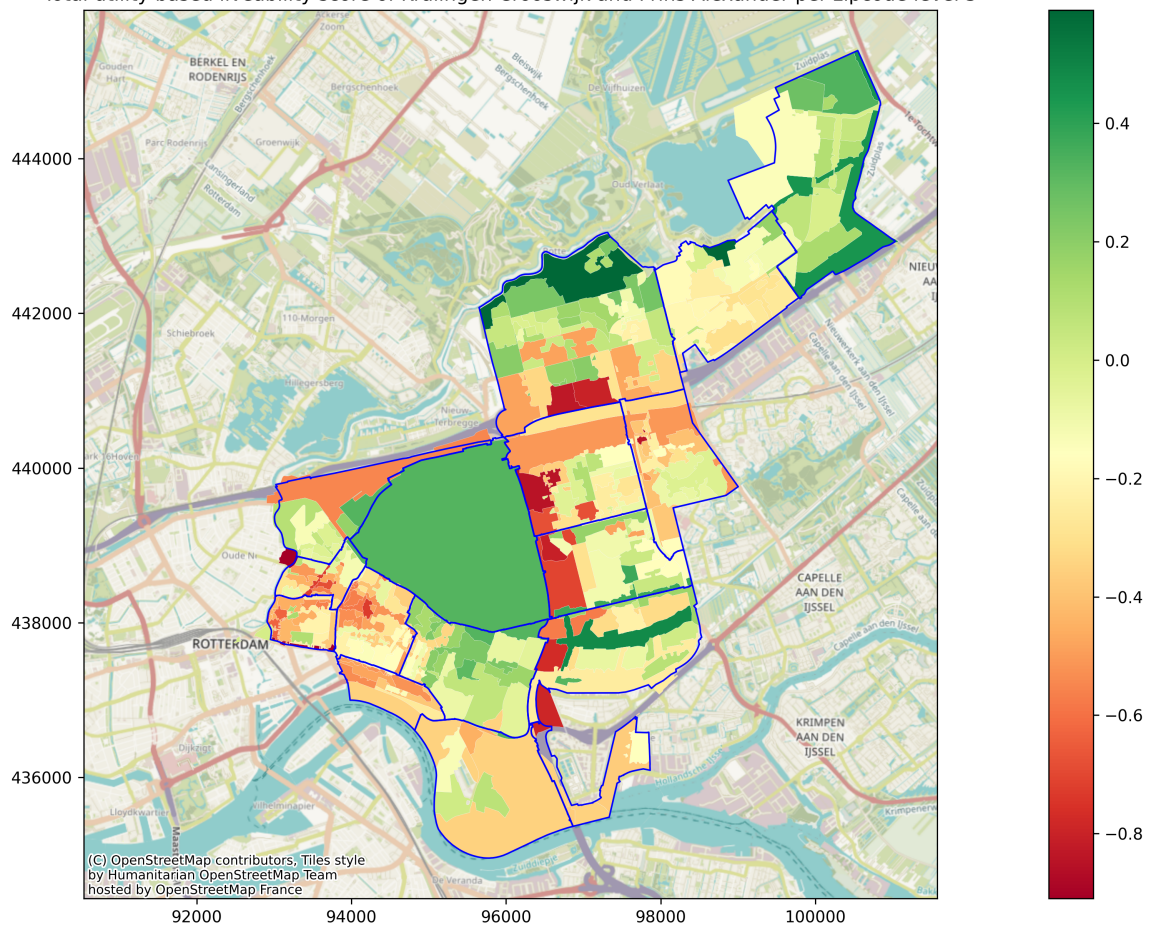


Figure E.1: Prins Alexander and Kralingen Crooswijk mapped together

F

Quantile plots of utility-based  
liveability score for Rotterdam



Figure F.1: Quantile plot of image utility score, number 2



Figure F.2: Quantile plot of image utility score, number 3

# G

## Street view images analysis of neighbourhoods

### G.1. Witte dorp



Figure G.1: Random street view images with their utility score from Witte Dorp



Figure G.2: Random street view images with their utility score from Witte Dorp

## G.2. Bospolder



Figure G.3: Random street view images with their utility score from Bospolder



Figure G.4: Random street view images with their utility score from Bospolder

### G.3. Tussendijken



Figure G.5: Random street view images with their utility score from Tussendijken



Figure G.6: Random street view images with their utility score from Tussendijken

### G.4. Nesselande

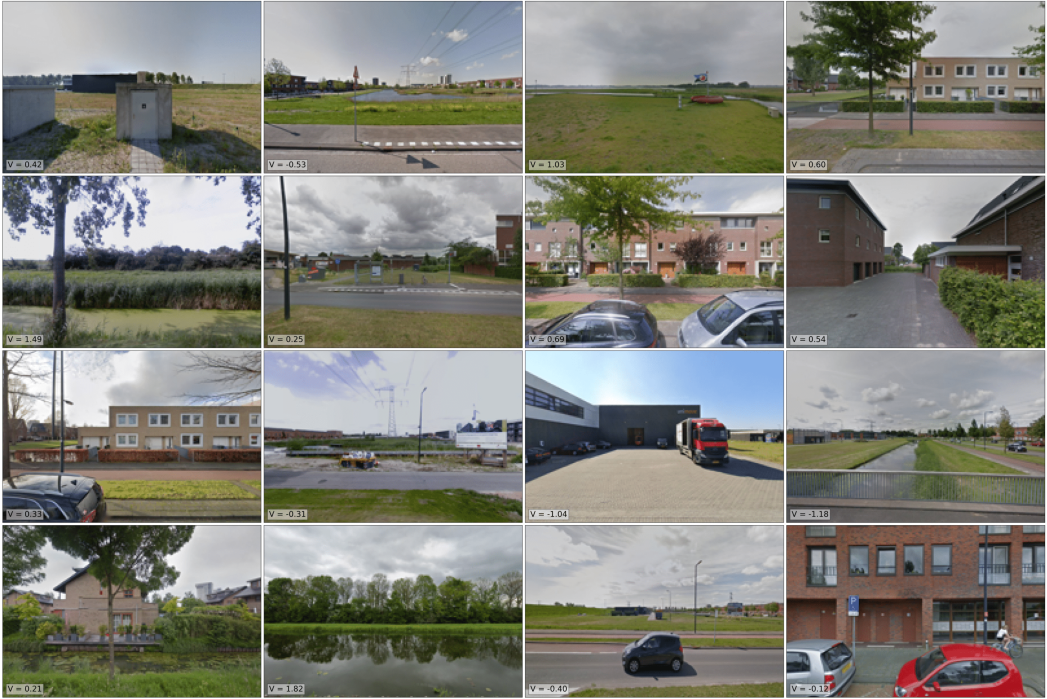


Figure G.7: Random street view images with their utility score from Nesselande





Figure G.8: Random street view images with their utility score from Nesselande

G.5. Het Lage Land



Figure G.9: Random street view images with their utility score from Het Lage Land



Figure G.10: Random street view images with their utility score from Het Lage Land

### G.6. Kralingseveer



Figure G.11: Random street view images with their utility score from Kralingseveer

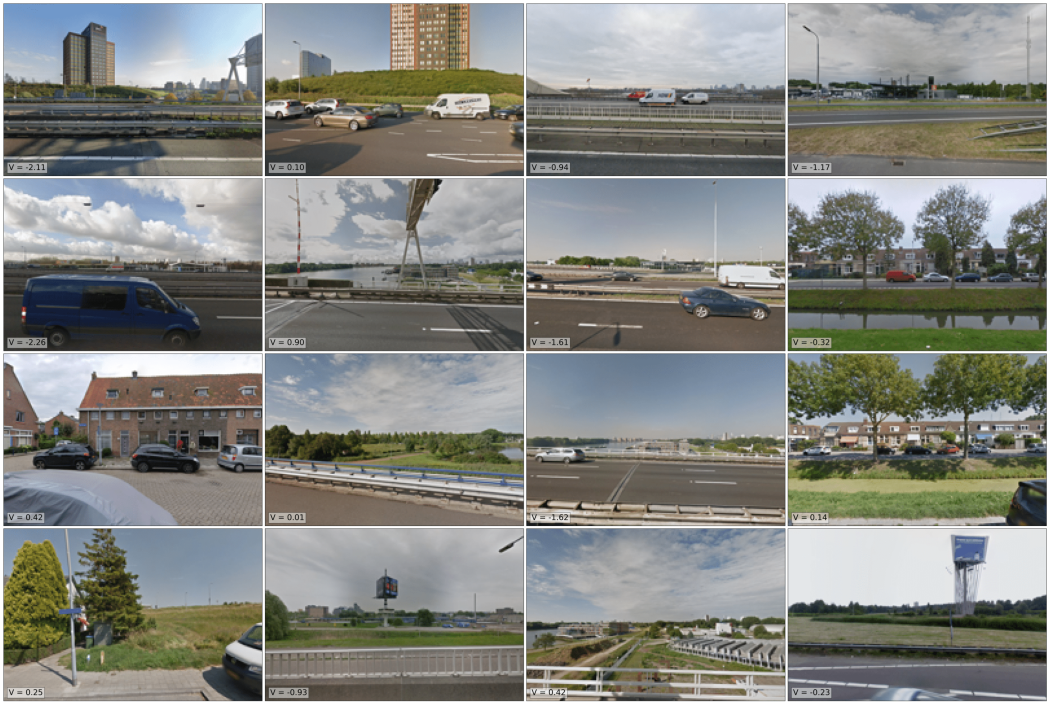


Figure G.12: Random street view images with their utility score from Kralingseveer

### G.7. Kralingse Bos



Figure G.13: Random street view images with their utility score from Kralingse Bos



Figure G.14: Random street view images with their utility score from Kralingse Bos

### G.8. Kralingen Oost



Figure G.15: Random street view images with their utility score from Kralingen Oost



Figure G.16: Random street view images with their utility score from Kralingen Oost

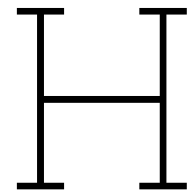
### G.9. Kralingseveer



Figure G.17: Random street view images with their utility score from Struisenburg



Figure G.18: Random street view images with their utility score from Struisenburg



## Additional analyzed segmentation algorithms

Compact Watershed is a variant of the traditional watershed algorithm, which draws inspiration from the concept of hydrological watershed basins. Initially, it computes the image gradient and identifies regional minima where markers are placed. These markers guide the watershed transform process to segment the image into catchment basins. The key feature of Compact Watershed is its "compactness" parameter, crucial for balancing between under-segmentation and over-segmentation. This parameter grants precise control over segmentation by adjusting how tightly regions adhere to object boundaries.

Unlike the traditional watershed method, which can lead to irregularly shaped and sized superpixels without boundary adherence, Compact Watershed improves this by enforcing compactness constraints on the resulting superpixels. This modification allows direct manipulation of the number and compactness of superpixels, addressing a significant limitation of the original watershed approach.

However, Compact Watershed may over-segment images, especially in areas with low gradients or noisy textures, and its effectiveness heavily relies on appropriate parameter tuning, particularly the compactness parameter. Despite these challenges, it remains efficient for segmenting images with distinct object boundaries, offering finer segmentation control compared to the conventional watershed algorithm (Vincent and Soille [136]).

Mean Shift is a non-parametric clustering algorithm widely used for image segmentation due to its robustness and ability to handle complex data distributions. Operating iteratively, Mean Shift moves each data point, representing pixels in image segmentation, towards the mode or peak of its underlying feature space distribution. This feature space typically encompasses colour and texture attributes, treating each pixel as a distinct data point.

In practical terms, Mean Shift identifies clusters by iteratively shifting pixels towards the densest regions within this feature space. This approach effectively segments the image into regions corresponding to these clusters without requiring prior knowledge of the number of segments. Moreover, Mean Shift demonstrates resilience against noise and cluttered backgrounds while accommodating irregularly shaped segments, making it suitable for various image segmentation tasks.

However, Mean Shift's computational demands can be significant, particularly when processing large images. Additionally, its performance is sensitive to parameter settings, notably the bandwidth parameter, which governs the size of the region used to estimate the density gradient. Proper parameter tuning is crucial to achieve optimal segmentation results (Comaniciu and Meer [19])

SLICO is an advanced version of the SLIC algorithm, designed to improve the boundary adherence of superpixels. It operates similarly to SLIC by clustering pixels based on color and spatial proximity

but introduces a term that penalizes deviations from image edges. This penalty term ensures that superpixels adhere more closely to object boundaries in the image. SLICO modifies the clustering process to favour superpixels that align with image boundaries. This results in better boundary adherence compared to the original SLIC algorithm. SLICO remains fast and efficient, suitable for real-time applications, and offers improved performance in maintaining the integrity of object boundaries within the image. However, it can still produce irregularly shaped superpixels in areas with complex textures or gradients. Additionally, while SLICO is sensitive to parameter settings, this sensitivity is less pronounced compared to some other algorithms (Achanta et al. [1]).

ERS (Efficient Graph-Based Image Segmentation) is an algorithm that segments images using a graph-based approach. It begins by constructing an over-segmented graph representation of the image, where each pixel is a node and edges represent relationships between pixels, such as proximity or colour similarity.

The algorithm then applies a graph-based clustering technique to merge similar regions while preserving boundaries. It iteratively merges adjacent regions based on a measure of dissimilarity until a stopping criterion is met. This process allows ERS to handle images with complex structures and textures, producing segments with well-defined boundaries.

ERS is parameterized by the desired number of segments, giving users control over the segmentation's granularity. However, it is relatively slower compared to some other segmentation algorithms, especially when processing large images. The performance of ERS is also sensitive to parameter settings, particularly the number of segments specified (Felzenszwalb and Huttenlocher [36]).



I

# Segmentations with Quickshift, Felzenszwalb and SLIC for the example image

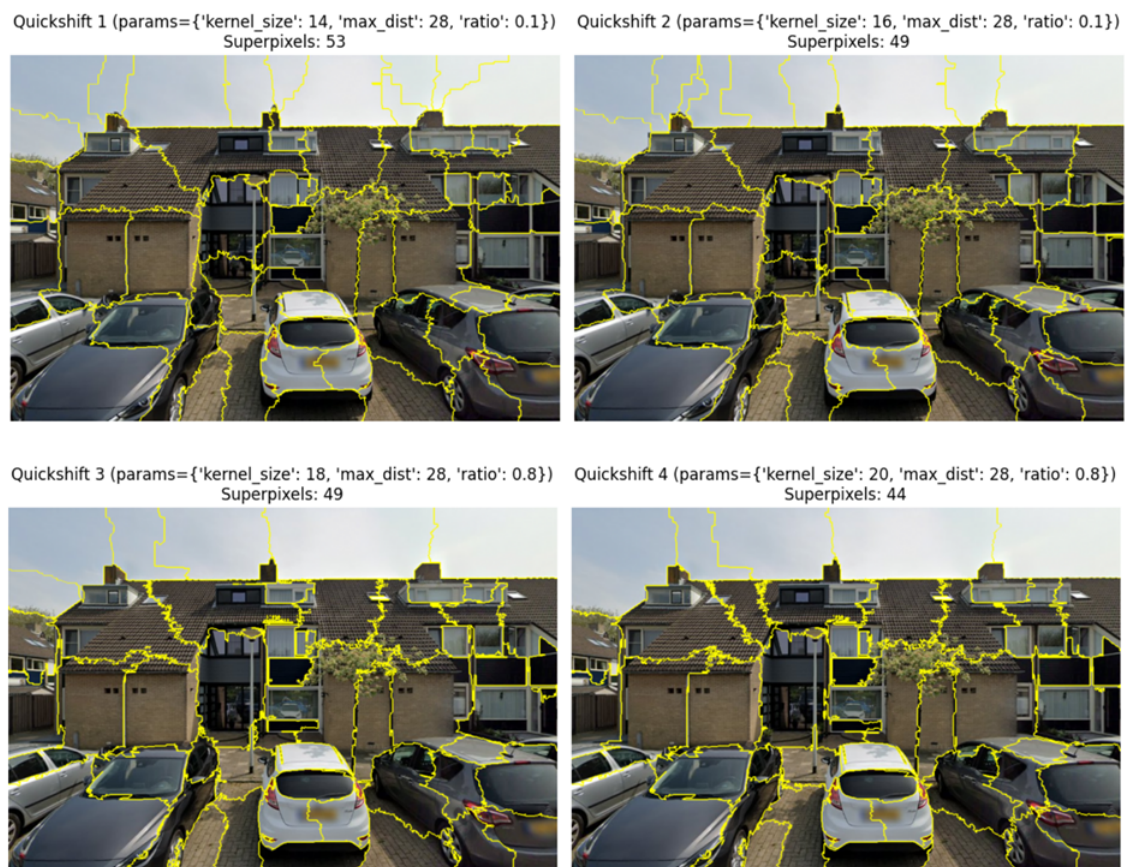


Figure I.1: Quickshift segmentations 1

Quickshift 5 (params={'kernel\_size': 14, 'max\_dist': 22, 'ratio': 0.5})  
Superpixels: 66



Quickshift 6 (params={'kernel\_size': 17, 'max\_dist': 26, 'ratio': 0.5})  
Superpixels: 46



Quickshift 7 (params={'kernel\_size': 20, 'max\_dist': 30, 'ratio': 0.5})  
Superpixels: 36



Quickshift 8 (params={'kernel\_size': 25, 'max\_dist': 34, 'ratio': 0.5})  
Superpixels: 23



**Figure I.2:** Quickshift segmentations 2

Felzenszwalb 1 (params={'scale': 900, 'sigma': 0.5, 'min\_size': 150}) Superpixels: 154  
Felzenszwalb 2 (params={'scale': 1500, 'sigma': 0.5, 'min\_size': 150}) Superpixels: 113



Felzenszwalb 3 (params={'scale': 2000, 'sigma': 0.5, 'min\_size': 150}) Superpixels: 87



Felzenszwalb 4 (params={'scale': 900, 'sigma': 0.5, 'min\_size': 50}) Superpixels: 297



Figure 1.3: Felzenszwalb segmentations 1

Felzenszwalb 5 (params={'scale': 1500, 'sigma': 0.5, 'min\_size': 50})  
Superpixels: 208



Felzenszwalb 6 (params={'scale': 2000, 'sigma': 0.5, 'min\_size': 50})  
Superpixels: 159



Felzenszwalb 7 (params={'scale': 1500, 'sigma': 0.5, 'min\_size': 100})  
Superpixels: 138



Felzenszwalb 8 (params={'scale': 2000, 'sigma': 0.5, 'min\_size': 100})  
Superpixels: 110



Figure I.4: Felzenszwalb segmentations 2

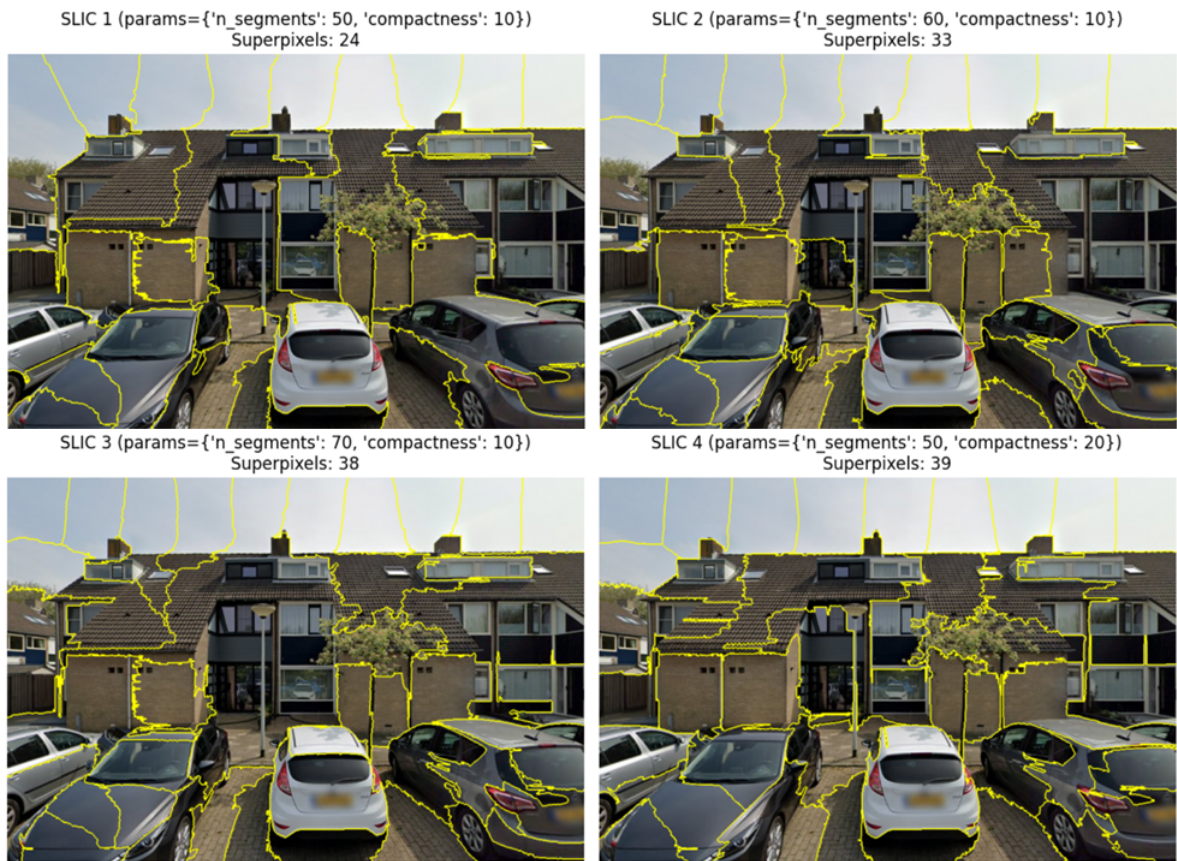


Figure I.5: SLIC segmentations 1

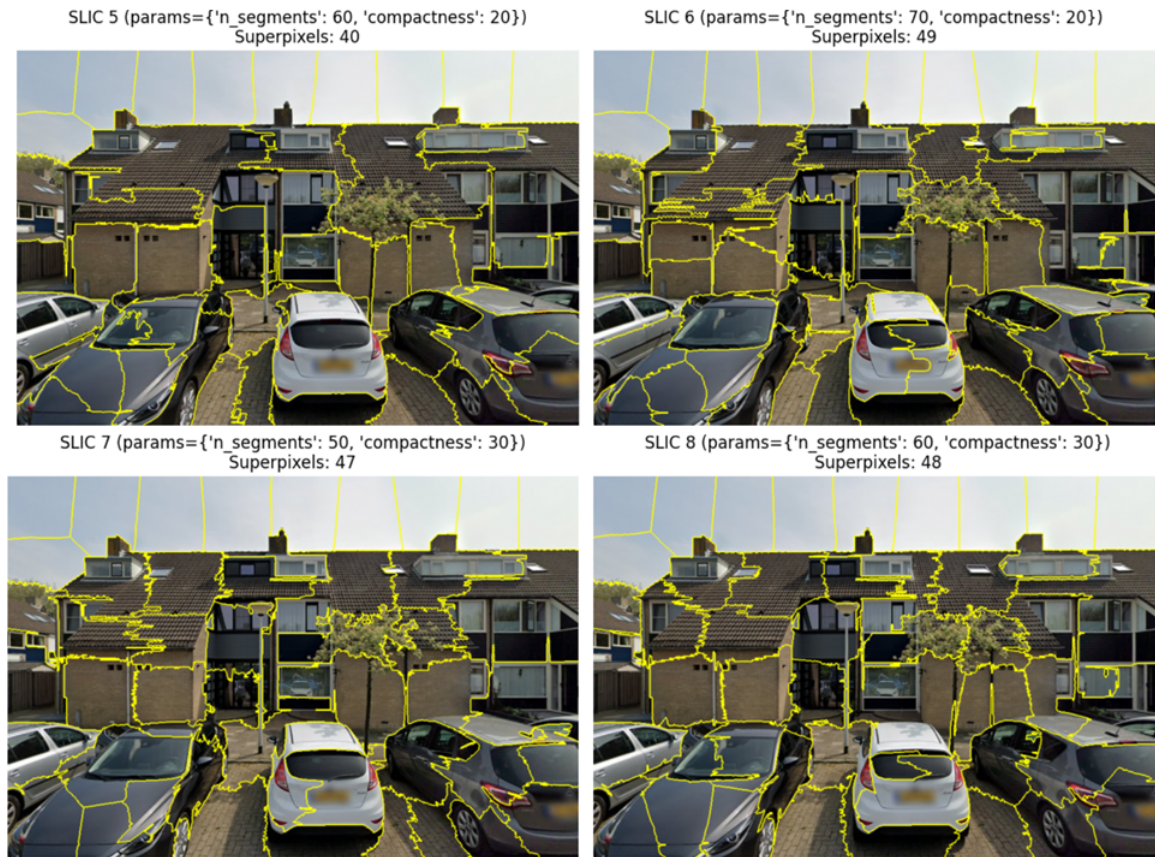


Figure I.6: SLIC segmentations 2

Segmentation	ARI	Jaccard	Dice	Vol
Quickshift 1	0.01	0.76	0.87	[5.21 0.48]
Quickshift 2	0.01	0.75	0.86	[5.08 0.50]
Quickshift 3	0.01	0.75	0.86	[5.08 0.48]
Quickshift 4	0.01	0.75	0.86	[4.88 0.49]
Quickshift 5	0.01	0.76	0.87	[5.41 0.46]
Quickshift 6	0.01	0.75	0.86	[5.04 0.48]
Quickshift 7	0.01	0.75	0.85	[4.69 0.50]
Quickshift 8	0.01	0.74	0.84	[4.24 0.54]
Felzenszwalb 1	0.04	0.61	0.76	[5.07 0.34]
Felzenszwalb 2	0.03	0.60	0.75	[5.07 0.34]
Felzenszwalb 3	0.02	0.60	0.75	[5.07 0.34]
Felzenszwalb 4	0.04	0.61	0.76	[5.07 0.34]
Felzenszwalb 5	0.03	0.60	0.76	[5.07 0.34]
Felzenszwalb 6	0.02	0.60	0.75	[5.07 0.34]
Felzenszwalb 7	0.03	0.60	0.75	[5.07 0.34]
Felzenszwalb 8	0.02	0.60	0.75	[5.07 0.34]
SLIC 1	0.01	0.79	0.88	[4.08 0.60]
SLIC 2	0.01	0.79	0.88	[4.33 0.51]
SLIC 3	0.01	0.79	0.88	[4.55 0.48]
SLIC 4	0.01	0.79	0.88	[4.89 0.51]
SLIC 5	0.01	0.79	0.88	[4.87 0.50]
SLIC 6	0.01	0.79	0.88	[5.15 0.49]
SLIC 7	0.01	0.79	0.88	[5.22 0.50]
SLIC 8	0.01	0.79	0.88	[5.28 0.50]

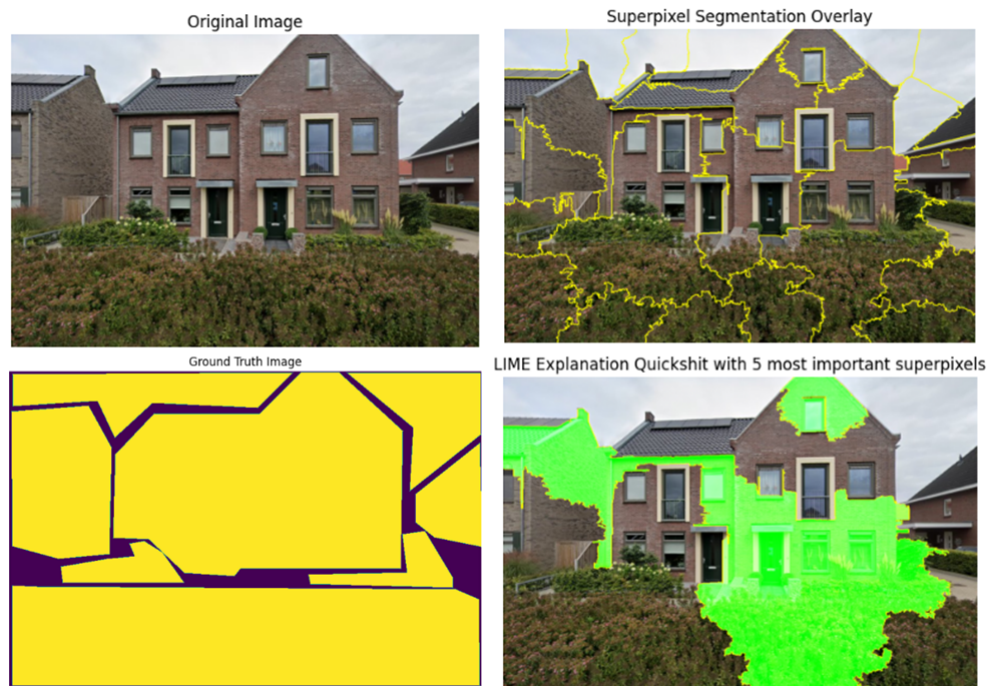
**Table I.1:** Ground Truth Metrics results

# J

## LIME case studies

### J.1. Case study 2

Figure J.1 shows the street view image, its segmentation, its ground truth image and the 5 most influential superpixels.



**Figure J.1:** Case 2: street view image, segmented image, ground truth image, 5 most influential superpixels

The results of the LIME explanation are given in table J.1.



Information	Score
Segmentation algorithm	SLIC(segments=50, compactness=10)
ARI	0.01
Jaccard	0.89
DICE	0.94
Vol	[4.14 , 0.40]
Mean Utility Score of perturbed images	1.21
Classification Threshold	1.18
Utility score original image	0.39
Classification original image	'bad'
BCR	0.48
Coefficient of Variation	2.80

Table J.1: Case study 2: lime analysis results

Figure J.2 shows the LIME with all superpixels effect, a heatmap of the superpixel weights, all positive superpixels and all negative superpixels.

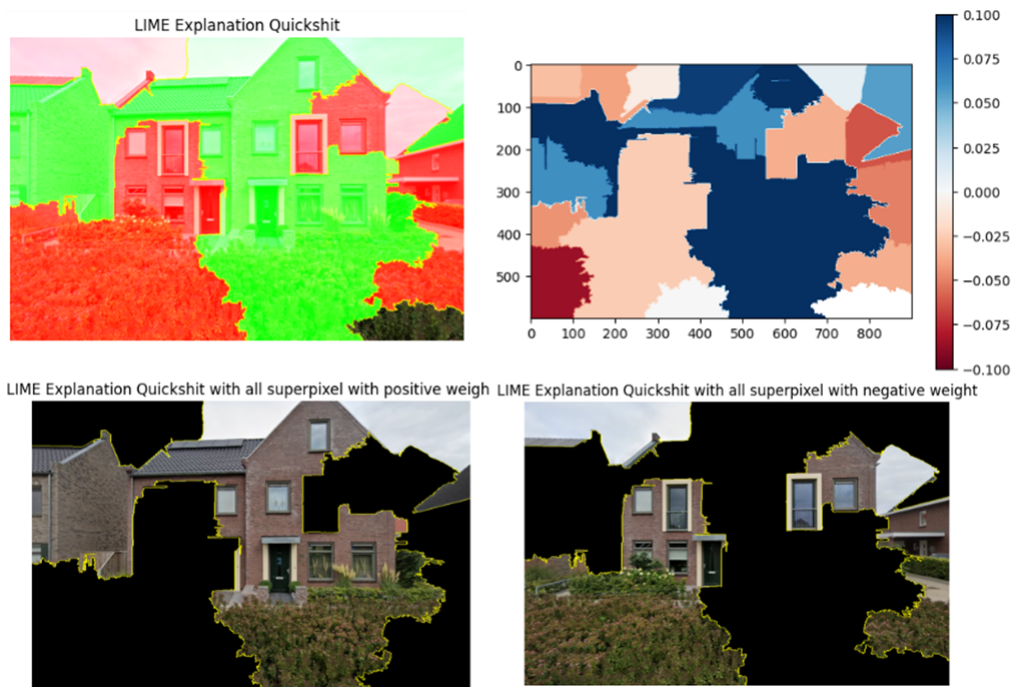
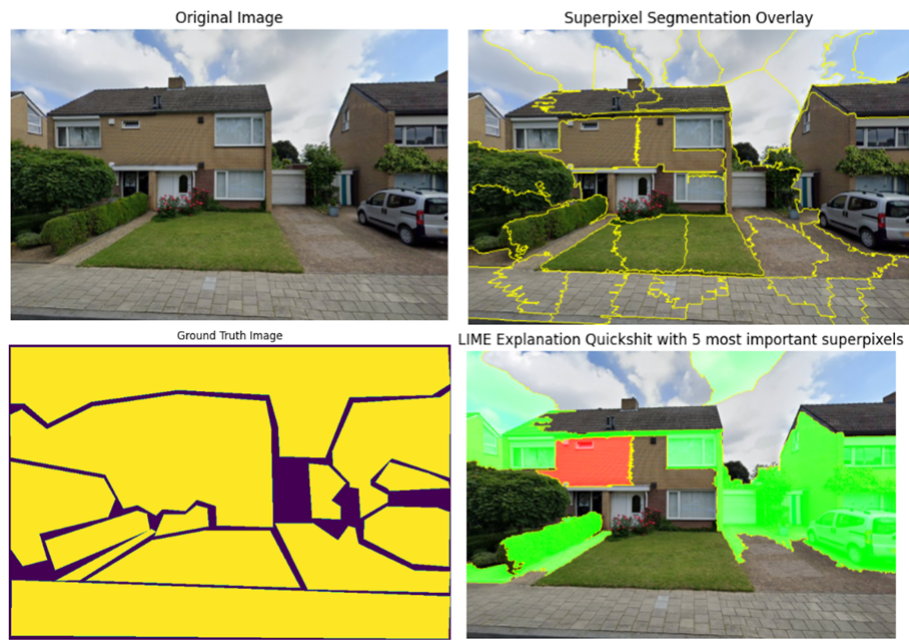


Figure J.2: Case 2: Lime explanation with all superpixels, superpixel weight heatmap, all positive superpixels, all negative superpixels

### J.2. Case study 3

Figure J.3 shows the street view image, its segmentation, its ground truth image and the 5 most influential superpixels.



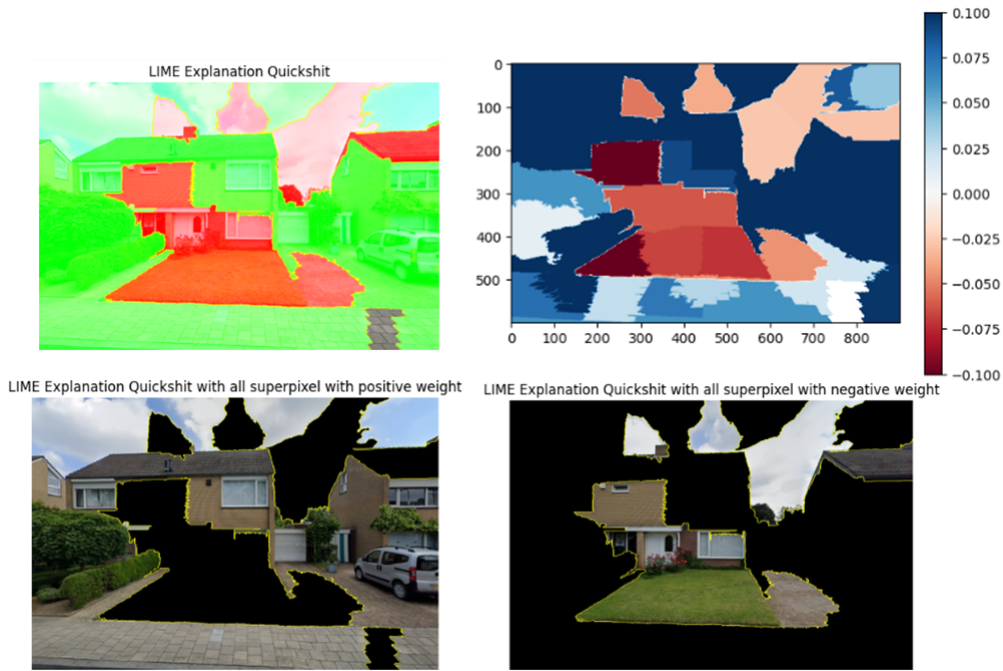
**Figure J.3:** Case 3: street view image, segmented image, ground truth image, 5 most influential superpixels

The results of the LIME explanation are given in table J.2.

Information	Score
Segmentation algorithm	SLIC(segments=50, compactness=10)
ARI	-0.01
Jaccard	0.87
DICE	0.93
Vol	[4.59 , 0.51]
Mean Utility Score of perturbed images	0.40
Classification Threshold	0.42
Utility score original image	-0.11
Classification original image	'bad'
BCR	0.52
Coefficient of Variation	2.79

**Table J.2:** Case study 3: lime analysis results

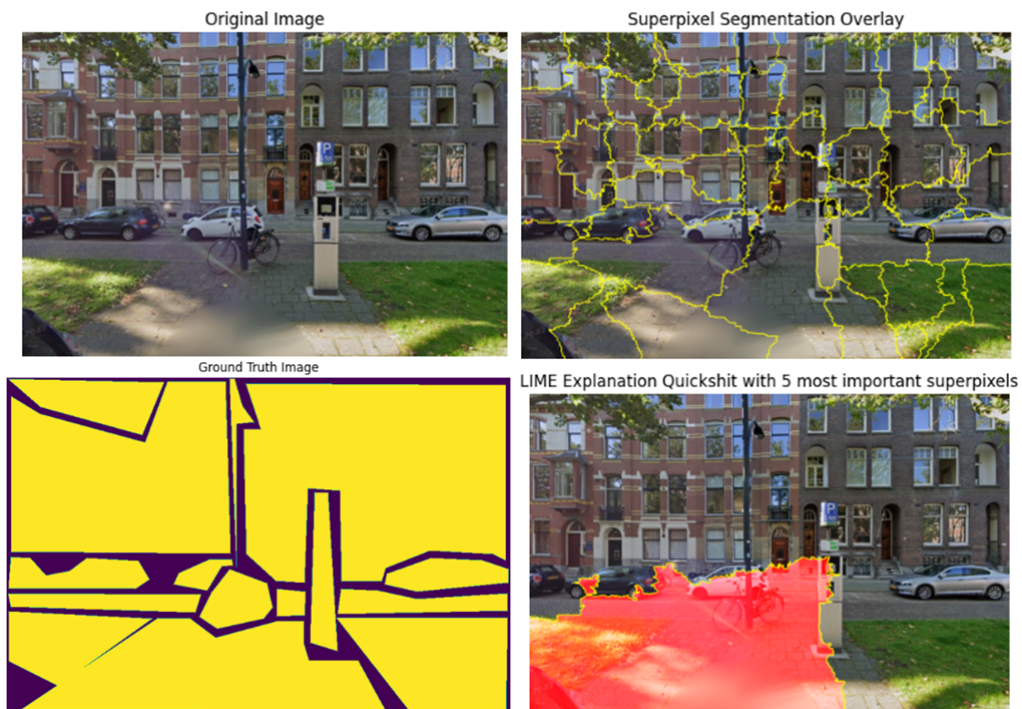
Figure J.4 shows the LIME with all superpixels effect, a heatmap of the superpixel weights, all positive superpixels and all negative superpixels.



**Figure J.4:** Case 3: Lime explanation with all superpixels, superpixel weight heatmap, all positive superpixels, all negative superpixels

### J.3. Case study 4

Figure J.5 shows the street view image, its segmentation, its ground truth image and the 5 most influential superpixels.



**Figure J.5:** Case 4: street view image, segmented image, ground truth image, 5 most influential superpixels

The results of the LIME explanation are given in table J.3.

Information	Score
Segmentation algorithm	Quickshift(kernel size=16, max dist=28, ratio=0.1)
ARI	-0.01
Jaccard	0.87
DICE	0.93
Vol	[4.59 , 0.51]
Mean Utility Score of perturbed images	0.62
Classification Threshold	0.6
Utility score original image	0.11
Classification original image	'bad'
BCR	0.46
Coefficient of Variation	-1.53

Table J.3: Case study 4: lime analysis results

Figure J.6 shows the LIME with all superpixels effect, a heatmap of the superpixel weights, all positive superpixels and all negative superpixels.

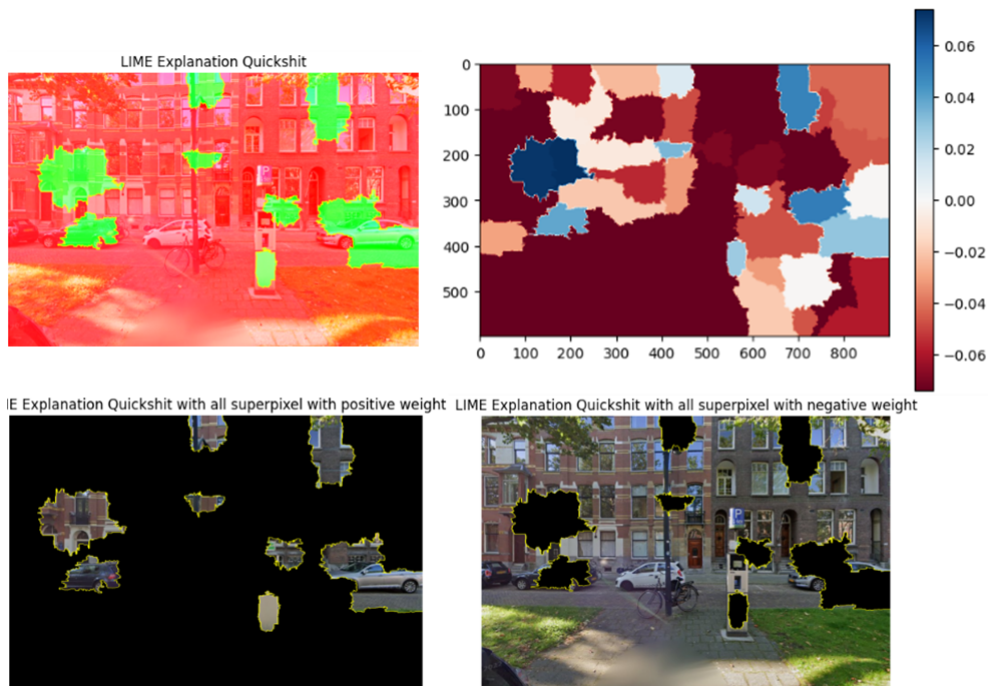
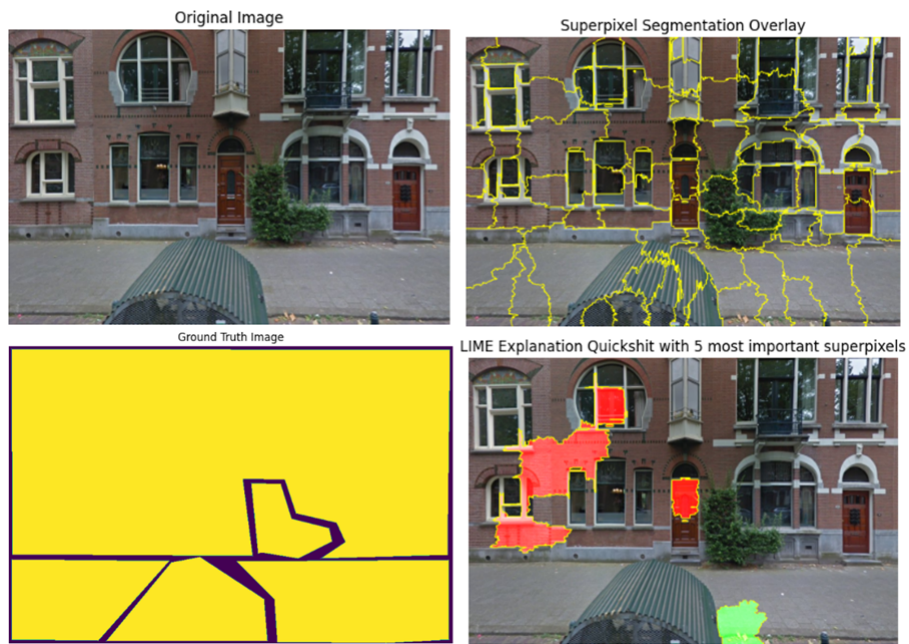


Figure J.6: Case 4: Lime explanation with all superpixels, superpixel weight heatmap, all positive superpixels, all negative superpixels

### J.4. Case study 5

Figure J.7 shows the street view image, its segmentation, its ground truth image and the 5 most influential superpixels.



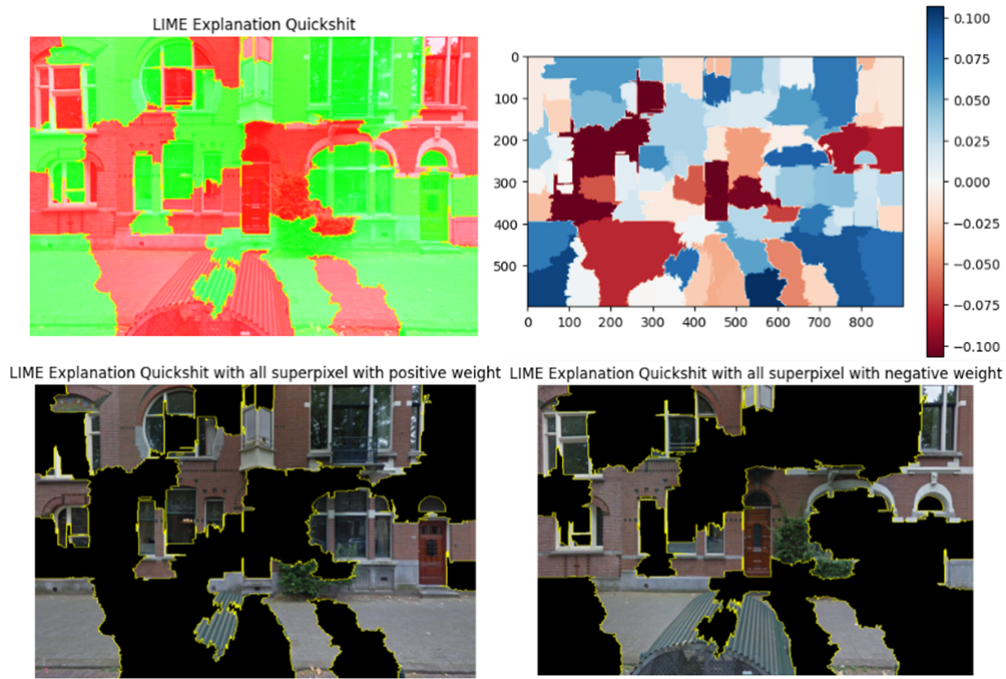
**Figure J.7:** Case 5: street view image, segmented image, ground truth image, 5 most influential superpixels

The results of the LIME explanation are given in table J.4.

Information	Score
Segmentation algorithm	Quickshift(kernel size=14, max dist=22, ratio=0.5)
ARI	0.01
Jaccard	0.92
DICE	0.96
Vol	[5.90 , 0.31]
Mean Utility Score of perturbed images	0.28
Classification Threshold	0.29
Utility score original image	-1.13
Classification original image	'bad'
BCR	0.5
Coefficient of Variation	64.90

**Table J.4:** Case study 5: lime analysis results

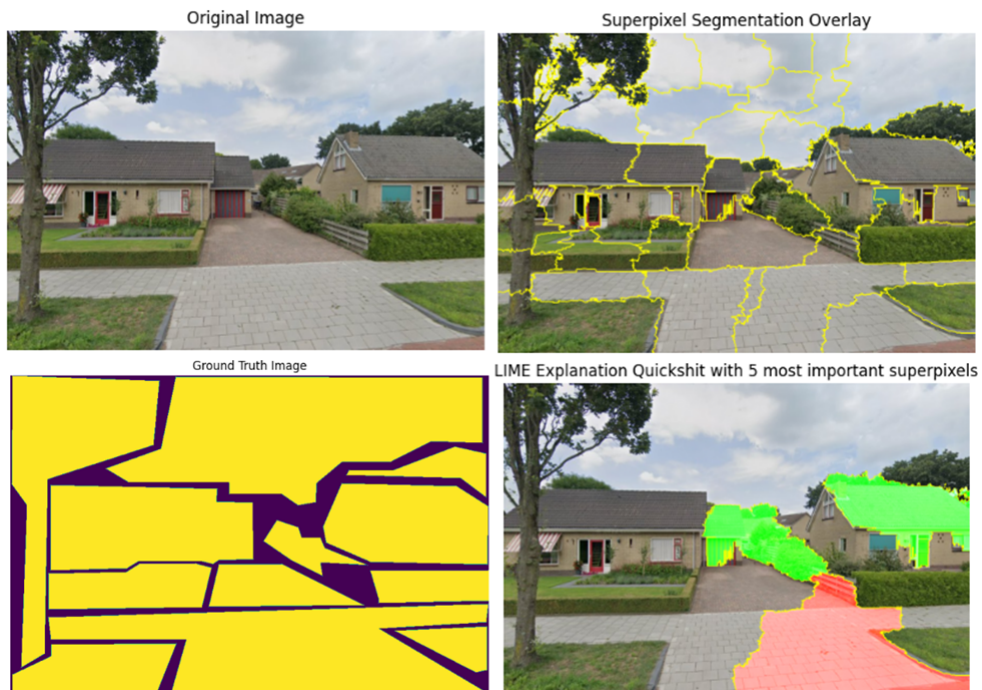
Figure J.8 shows the LIME with all superpixels effect, a heatmap of the superpixel weights, all positive superpixels and all negative superpixels.



**Figure J.8:** Case 5: Lime explanation with all superpixels, superpixel weight heatmap, all positive superpixels, all negative superpixels

### J.5. Case study 6

Figure J.9 shows the street view image, its segmentation, its ground truth image and the 5 most influential superpixels.



**Figure J.9:** Case 6: street view image, segmented image, ground truth image, 5 most influential superpixels

The results of the LIME explanation are given in table J.5.

Information	Score
Segmentation algorithm	Quickshift(kernel size=18, max dist=28, ratio=0.8)
ARI	0.01
Jaccard	0.83
DICE	0.91
Vol	[4.97 , 0.60]
Mean Utility Score of perturbed images	1.10
Classification Threshold	1.09
Utility score original image	1.15
Classification original image	'good'
BCR	0.51
Coefficient of Variation	7.96

Table J.5: Case study 6: lime analysis results

Figure J.10 shows the LIME with all superpixels effect, a heatmap of the superpixel weights, all positive superpixels and all negative superpixels.

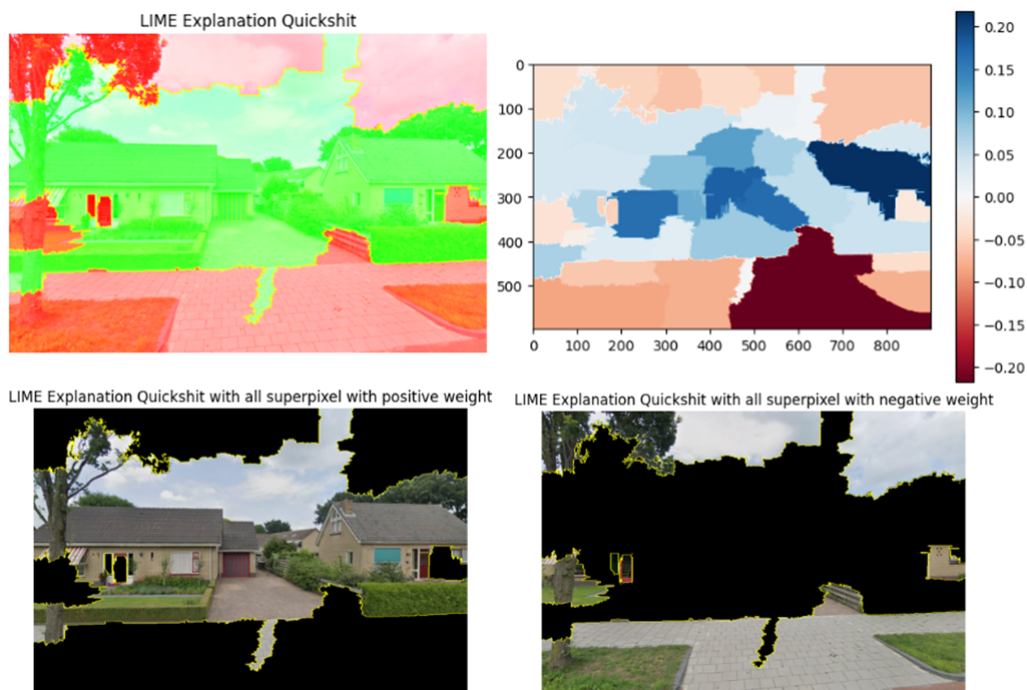
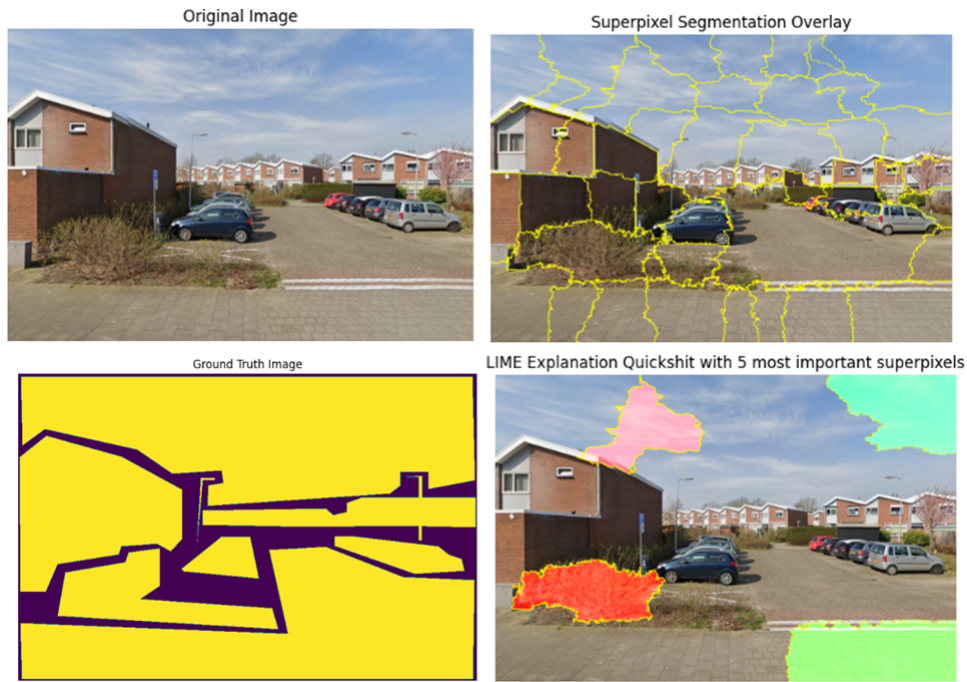


Figure J.10: Case 6: Lime explanation with all superpixels, superpixel weight heatmap, all positive superpixels, all negative superpixels

### J.6. Case study 7

Figure J.11 shows the street view image, its segmentation, its ground truth image and the 5 most influential superpixels.



**Figure J.11:** Case 7: street view image, segmented image, ground truth image, 5 most influential superpixels

The results of the LIME explanation are given in table J.6.

Information	Score
Segmentation algorithm	Quickshift(kernel size=17, max dist=26, ratio=0.5)
ARI	0.01
Jaccard	0.83
DICE	0.91
Vol	[4.97 , 0.60]
Mean Utility Score of perturbed images	1.50
Classification Threshold	1.47
Utility score original image	-0.26
Classification original image	'bad'
BCR	0.46
Coefficient of Variation	20.37

**Table J.6:** Case study 7: lime analysis results

Figure J.12 shows the LIME with all superpixels effect, a heatmap of the superpixel weights, all positive superpixels and all negative superpixels.



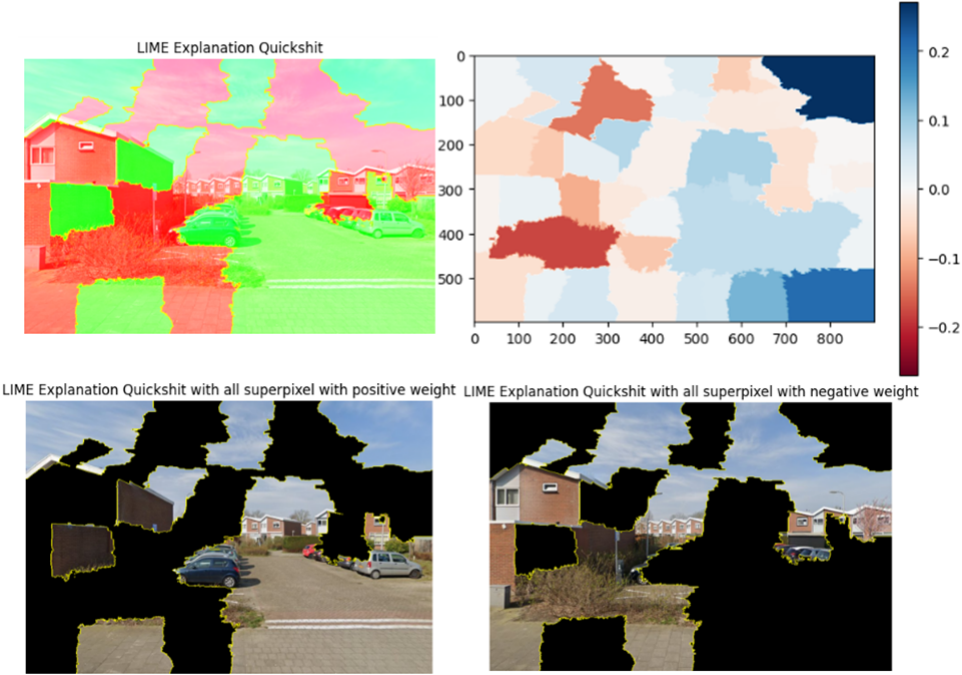


Figure J.12: Case 7: Lime explanation with all superpixels, superpixel weight heatmap, all positive superpixels, all negative superpixels