

Student thesis series INES nr 726

Street-level greenery mapping and urban heat modelling using deep learning and street view imagery

Toon Lambrecht

2025
Department of
Physical Geography and Ecosystem Science
Lund University
Sölvegatan 12
S-223 62 Lund
Sweden



Toon Lambrecht (2025)

Street-level greenery mapping and urban heat modelling using deep learning and street view imagery

Master degree thesis, 30 credits in *GIS and Remote Sensing*

Department of Physical Geography and Ecosystem Science, Lund University

Level: Master of Science (MSc)

Course duration: *January 2025 until June 2025*

Disclaimer

This document describes work undertaken as part of a program of study at the University of Lund. All views and opinions expressed herein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

Street-level greenery mapping and urban heat modelling using deep learning and street view imagery

Toon Lambrecht

Master thesis, 30 credits, in *GIS and Remote Sensing*

Supervisor 1:

Rachid Oucheikh, Lund University

Supervisor 2:

Pengxiang Zhao, Lund University

Exam committee:

Helena Elvén Eriksson, Lund University

Karin Hall, Lund University

Acknowledgements

I would like to thank both of my supervisors, Rachid Oucheikh and Pengxiang Zhao, for their support throughout this project. Inexhaustibly, they provided extensive feedback and taught me many things within the field of geoinformatics. I would also like to thank Frederik Wienker, who was my thesis opponent and shared many thoughtful comments regarding the report. Special thanks to Bart and Chrissi, who voluntarily read my thesis and provided additional feedback as external persons and helped me structuring the report. Thanks also to Joseph for all the much needed and fun study breaks during the thesis period. Lastly, I want to thank my mother, Ranja, and my partner, Josefien, for their love, support, and encouragement throughout my studies and during the writing of this thesis.

Abstract

In an urbanizing world, challenged by climate change and rising temperatures, street vegetation can help mitigate urban temperatures. However, scaling the understanding of street vegetation's cooling effects from local observations to the city level introduces challenges. While several studies have assessed the influence of street greenery on land surface temperature (LST) at the city scale using street view (SV) data, few have distinguished between different types of greenery. Additionally, the combined investigation of spatial non-stationarity and non-linear trends remains underexplored. This study fine-tuned a semantic segmentation model for the city of Zurich using SV imagery. The model extracted urban features - trees, shrubs, grass, buildings, and sky - by calculating view factors (VFs). Street orientation and water presence were quantified using ancillary data. These urban characteristics served as inputs for a Geographically Weighted Random Forest (GWRF) model, complemented by SHAP analysis, to evaluate their impact on LST. Moreover, different approaches to quantifying the extracted features via VFs were assessed.

The fine-tuned deep learning model effectively captured different urban features from the SV data, achieving high Intersection over Union (IoU) scores across the extracted classes: tree (IoU = 0.91), shrub (IoU = 0.76), grass (IoU = 0.83), building (IoU = 0.92), and sky (IoU = 0.99). The GWRF model proved effective at predicting LST, with good evaluation metrics for both training ($R^2 = 0.89$, MAPE = 2.10%) and test data ($R^2 = 0.87$, MAPE = 2.26%). SHAP analysis revealed that buildings had the strongest positive effect on LST at moderate to high VFs, with a clear non-linear trend. Conversely, trees had the strongest cooling effect under similar VF ranges. Shrubs had a moderate, yet more spatially distributed cooling effect, suggesting their potential in areas where space for large canopy trees is limited.

The use of a modified, fine-tuned network proved effective in improving classification results of the street view images. In future research, this approach can be extended with additional classes to investigate the influence of specific urban characteristics. The applied method, integrating GWRF and SHAP analysis, showed potential for studying the relationship with LST while considering spatial non-stationarity and non-linear effects. By applying this method to other case studies, its wider applicability in analysing the dynamics of LST can be evaluated. Such studies could reveal spatial trends and support urban planners and policymakers in effectively mitigating LST in urban environments.

Keywords: Physical Geography and Ecosystem Analysis; Google Street View, Semantic Segmentation, Deep Learning, Street Greenery, Urban Heat Modelling, Geographically Weighted Random Forest., Zurich.

Table of Contents

1. Introduction	1
1.1. Study aim.....	2
1.2. Study scope	2
1.3. Study outline	2
2. Background.....	3
2.1. Urban heat island effect.....	3
2.2. Urban cooling through plants	4
2.3. Review of SV data applications in temperature modelling.....	5
2.4. Semantic segmentation.....	7
2.5. Machine learning.....	9
3. Method and Materials.....	10
3.1. Study area and data collection.....	11
3.2. Semantic segmentation.....	16
3.3. View factor calculations	21
3.4. Exploring the relationships between urban features and LST.....	24
4. Results	25
4.1. Semantic segmentation.....	25
4.2. View factor calculations	27
4.3. Exploring the relationships between urban features and LST.....	31
5. Discussion	37
5.1. Data collection.....	37
5.2. Semantic segmentation.....	38
5.3. View factor calculations	39
5.4. Exploring the relationships between urban features and LST.....	40
5.5. Future research	44
6. Conclusions	44
References	46
Appendices	i
Appendix A: Availability of Google Street View images in Zurich	i
Appendix B: Evaluation metrics	ii
Appendix C: Individual maps of distribution of model variables	iv
Appendix D: SHAP values from GWRF for test data.....	xi
Appendix E: Individual maps of SHAP values.....	xii

Glossary of terms

Abbreviation	Name	Definition
Adam	Adaptive Moment Estimation	An optimization algorithm combining momentum and adaptive learning rates for training neural networks.
API	Application Programming Interface	A set of protocols enabling different software systems to communicate.
COCO	Common Objects in Context	A large-scale dataset of everyday scene images annotated for object detection, segmentation, and captioning.
CPU	Central Processing Unit	The primary processor of a computer.
EW	East-West	East-West oriented streets
FOV	Field Of View	The angular area visible for a GSV image.
GEE	Google Earth Engine	A cloud-based platform for large-scale geospatial data analysis.
GPU	Graphics Processing Unit	A specialized processor optimized for parallel computations.
GSV	Google Street View	A Google Maps/Earth feature providing interactive 360° panoramic street-level imagery.
GVI	Green View Index	A metric quantifying the proportion of greenery visible in street-level imagery.
GWR	Geographically Weighted Regression	A linear regression technique that accounts for spatial non-stationarity.
GWRF	Geographically Weighted Random Forest	A variant of random forest that adjusts model parameters based on geographic location.
HTTPS	Hypertext Transfer Protocol Secure	A secure version of HTTP employing encryption via TLS/SSL.
IoU	Intersection over Union	A metric measuring the overlap between predicted and ground-truth bounding boxes in object detection.
JSON	JavaScript Object Notation	A lightweight data-interchange format using human-readable text.
LST	Land Surface Temperature	The temperature of the Earth's surface measured via satellite remote sensing.
MAE	Mean Absolute Error	The average of absolute differences between predicted and actual values in a dataset.
MAPE	Mean Absolute Percentage Error	The average of absolute percentage differences between forecasts and actual values.
MRT	Mean Radiant Temperature	The uniform temperature of an imaginary enclosure where radiant exchange equals that of the actual environment.
MSE	Mean Squared Error	The average of squared differences between predicted and observed values.
MVC	Multi-View Collection	Images captured from multiple viewing angles to describe the hemisphere.

NDBI	Normalized Difference Building Index	A remote sensing index for estimating built-up areas.
NDVI	Normalized Difference Vegetation Index	A remote sensing index for estimating vegetation.
NDWI	Normalized Difference Water Index	A remote sensing index for estimating water areas.
NS	North-South	North-South oriented streets
OOB	Out-Of-Bag	A cross-validation method in random forests using data not sampled during bootstrap aggregation.
P	Precision	The ratio of correctly predicted positive observations to the total predicted positives.
PET	Physiological Equivalent Temperature	An index evaluating perceived human thermal comfort based on environmental variables.
PIC	Panoramic Image collection	A set of images captured in a full 360° panoramic view of a scene.
PSPNet	Pyramid Scene Parsing Network	A deep learning architecture for semantic segmentation using multi-scale contextual pooling.
R	Recall	The ratio of correctly predicted positive observations to all actual positives.
ResNet	Residual Network	A convolutional neural network design utilizing skip connections to ease training of deep layers.
SHAP	Shapley Additive Explanations	A method that explains model predictions by attributing contributions of each feature.
SV	Street View	Generic term for street-level panoramic imagery.
SVF	Sky View Factor	The fraction of visible sky in a given outdoor location.
TLM	Topographic Landscape Model	A spatial model representing terrain features and topography for Switzerland.
UHI	Urban Heat Island	The phenomenon where urban areas experience higher temperatures than surrounding rural areas.
UMT	Urban Morphology Types	Categories classifying urban form based on patterns such as building layout and street design.
VF	View Factor	Metric for quantifying the relative presence of a certain feature in an image.
VPD	Vapour Pressure Deficit	The difference between the amount of moisture in the air and its maximum at saturation.

1. Introduction

Projections indicate that by 2050, 68% of the world's population will live in urban areas (United Nations (UN), 2019). While urbanization patterns and the associated challenges vary across continents and regions, the need for coordinated urban policies to address these difficulties is universal (Kundu & Pandey, 2020). One of these major challenges in high-density areas is their vulnerability to heat waves, which are intensified in urban environments due to built-up, impervious surfaces. These surfaces absorb large amounts of solar radiation and gradually release it as longwave radiation, contributing to local climate warming (Depietri et al., 2011).

Green spaces offer solutions to this. Plants help regulate air temperature through evapotranspiration, absorption of photosynthetically active radiation, and reflection of near-infrared radiation (Depietri et al., 2011). However, despite the documented benefits of green spaces regarding temperature regulation, shifting the focus to the role of street greenery at a city scale, presents new challenges. Measuring its impact within the broader urban fabric is complex, as land surface temperature (LST) is influenced by multiple factors such as street geometry, surface permeability and proximity to large water bodies. Furthermore, studies that have been conducted often generalise vegetation into one category, ignoring the multitude of vegetation types within the streetscape (Naeem et al., 2018; Wang et al., 2023).

Traditionally, greenery is quantified using large-scale satellite data, which relies on vegetation indices such as the Normalized Difference Vegetation Index (NDVI), or through field surveys. While field surveys provide detailed and highly accurate information, they lack scalability. Satellite data, on the other hand, offers scalability but captures greenery only from a bird's-eye view, potentially missing understory vegetation such as grass, shrubs, or façade greenery (Lu et al., 2023; Zhang et al., 2019). As a result, relying solely on satellite data may not accurately reflect the 3D structure of the urban environment, making it difficult to distinguish between different vegetation types.

Street View (SV) data, widely available in many parts of the world, has emerged as a new method for studying and quantifying street greenery. Using SV data, the urban environment can be characterized from a pedestrian perspective, offering a better estimate of the 3D structure of the urban environment. This allows for fine-scale mapping of street greenery, capturing multiple layers and details that may not be visible from a bird's eye view. While SV images are currently limited to RGB bands, advances in image segmentation architectures have facilitated accurate greenery mapping for large datasets, making SV data suitable for city-scale applications (Lu et al., 2023; Ye et al., 2018).

Currently, studies that have used SV data to investigate the relationship between LST and greenery have made limited distinctions between different types of greenery or have included no other urban characteristics beyond greenery in their analyses. Furthermore, existing research on the relationship with LST has primarily relied on correlation analysis, global linear regression models, or geographically weighted regression (GWR) (Li et al., 2022; Wei et al., 2024; Zhang et al., 2019). While GWR can account for spatially non-stationary effects, it cannot capture non-linear or non-monotonic relationships (Subedi et al., 2018). Another limitation that in SV research that has received limited attention is the quantification of extracted features from deep learning networks. This is typically done through view factors (VFs), which describe the

relative presence of a specific feature in an SV image. To date, three common approaches have been used to calculate VFs: using individual images, panoramic images, and fisheye images (Li & Ratti, 2018; Wei et al., 2024; X. Li et al., 2015, Lu et al., 2023). However, no comparative study has yet been conducted to evaluate how these VF calculation methods affect the modelling of their relationship with LST.

1.1. Study aim

To better understand the role of street greenery in reducing LST at a city scale, this study aims at including different elements of the cityscape, while also distinguishing between types of greenery. To study fine scale influences, high resolution LST data is needed, which can be obtained through data upscaling. Additionally, a modelling approach is needed that can account for potential non-linear relationships and spatial non-stationarity. Furthermore, the study aims to assess how the quantification methods of VFs influence model behaviour, to identify potential limitations. To address these needs, the following research questions were formulated:

- 1) Can the urban environment with different types of street greenery, including grass, shrubs and trees, be accurately mapped using SV data through deep learning techniques?
- 2) What are the differences between different VF calculation methods and how do they influence model behaviour and performance?
- 3) What is the impact of different urban features on LST at the street level?

To address the research questions and aims, urban characteristics were derived from Google Street View (GSV) imagery using a custom fine-tuned segmentation network. The extracted features included built elements, three types of greenery (trees, shrubs, and grass) and the sky. Additionally, street orientation and water presence were quantified using ancillary data. The different methods for quantifying VFs were compared in terms of their consistency and impact on model performance. A modelling framework combining geographically weighted random forest (GWRF) and SHapley Additive exPlanations (SHAP) analysis was proposed to explore relationships between the extracted features and LST.

1.2. Study scope

This study focused on a single case study: the city of Zurich. It did not include comparisons across different urban environments or semantic segmentation models. Instead, the emphasis was on fine-tuning one model and selecting relevant semantic classes for urban characterization. To examine the relationship between urban features and LST, the workflow combining GWRF and SHAP analysis was implemented. While the workflow's application and limitations were explored, no direct comparison with other modelling techniques, such as GWR, was made.

1.3. Study outline

The structure of this thesis is as follows: the theoretical background reviews concepts related to urban heat islands (UHI), vegetation-based cooling, and the use of SV data for heat modelling. The methodology section covers the different steps undertaken to answer the research questions and the analysis outcomes are presented in the results section. The discussion section provides more context to the results, explicitly answers the research questions, addresses limitations, and suggests directions for future research.

2. Background

This research covers a broad range of topics, from plant physiology to machine learning. This section provides a brief overview of prior studies on UHIs, temperature mitigation through vegetation, key findings of related research, and remaining research gaps. Additionally, the methodological framework includes two machine learning techniques - semantic segmentation and GWRP - which are explained in more detail in this section.

2.1. Urban heat island effect

The UHI effect refers to the phenomenon where urban environments experience higher temperatures than their surrounding rural areas. This occurs due to urbanization and human activities that alter the local energy balance (Mohajerani et al., 2017; Santamouris, 2011). Urbanization contributes to heating primarily through the physical properties of urban materials, city geometry, and urban sprawl. Streets, pavements, and buildings often have a lower albedo, meaning they absorb more short- and longwave radiation and subsequently re-emit it as longwave radiation, increasing air temperatures. Additionally, these surfaces have a higher specific heat capacity, allowing them to store more thermal energy than natural landscapes and release it gradually, leading to sustained warming, especially at night (Aletba et al., 2021; Depietri et al., 2011; Mohajerani et al., 2017).

Street canyons - narrow corridors flanked by tall buildings - trap heat and restrict airflow. Their radiative geometry lowers the system's effective albedo due to multiple reflections of shortwave radiation between surfaces, increasing heat absorption. Furthermore, heat exchange between buildings reduces longwave radiation loss, while limited sky exposure hinders heat dissipation. The turbulent transfer of heat within street canyons further restricts cooling, preventing efficient heat escape into the atmosphere (Oke, 1982; Santamouris, 2011).

Beyond urban design, anthropogenic factors also contribute to city warming. Air conditioning systems release waste heat into the urban environment, worsening temperature increases. This creates a feedback loop, as rising temperatures drive greater cooling demand, which in turn generates more waste heat (Ghaddar et al., 2017; Lee et al., 2013). Ghaddar et al. (2017) found that air conditioning waste heat can increase street temperatures by 0.5°C during the day and up to 4.7°C at night in Beirut. Other sources of anthropogenic heat include human metabolism, vehicles, industry, and power plants. Although the relative contribution of each factor varies depending on a city's geographic structure, their combined effects typically result in two peaks of heat dissipation - one in the morning and another in the afternoon (Lee et al., 2013; Sailor, 2010).

Multiple studies have quantified the extent to which these factors contribute to urban heating. Takebayashi and Moriyama (2012) found that pavements can reach temperatures up to 20°C higher than grass during the day and remain up to 4°C warmer at night. Similarly, Le et al. (2020) observed that dense asphalt concrete in Hanoi can reach temperatures of up to 60°C on hot summer days.

Regarding air temperature, Sayad et al. (2024) identified a temperature difference of up to 6.9°C between urbanized areas in Annaba, Algeria, and their rural surroundings, with the mean daily difference ranging between 4.5°C and 4.8°C. Fahmy et al. (2023) analysed land use and land cover changes in Egypt's Sharqiyah Governorate from 2001 to 2022 using satellite data. Their

findings indicate that, over this period, LST increased by an average of 5°C in urban areas, compared to 3.98°C for the entire region, driven by urban expansion and declining vegetation both within and beyond city boundaries. Koomen and Diogo (2015) examined the Amsterdam region in the Netherlands, finding that the UHI effect leads to maximum air temperature differences of over 3°C compared to the surrounding countryside on moderately warm summer days, rising to as much as 5°C on warm days.

Although the magnitude of the UHI effect varies significantly depending on urban morphology, surface materials, and climate zone, all these studies point to a consistent trend: cities experience significantly higher temperatures than their surrounding areas. Urban vegetation offers a solution to this problem by providing several ecosystem services, one of the most important of which is temperature regulation (Depietri et al., 2011).

2.2. Urban cooling through plants

Plants contribute to urban cooling primarily through the reflection of solar radiation and evapotranspiration. Trees have a higher albedo than surrounding urban surfaces such as building facades, roofs, and pavements. They reflect a portion of incoming radiation, particularly in the infrared (IR) spectrum, reducing heat absorption by urban surfaces and creating shade (Deng et al., 2019). On top of reflectance, trees also absorb incoming solar radiation, lowering the absorbed radiation by other surfaces, and subsequently lowering ambient temperatures. The effectiveness of these cooling mechanisms - reflectance and shading - varies depending on tree species, leaf structure, canopy density, and urban morphology (Deng et al., 2019; Wu et al., 2024).

In addition to reflectance and absorption, trees contribute to urban cooling through evapotranspiration. Evapotranspiration is a cooling mechanism through which trees lower temperatures by transforming absorbed heat into latent heat. This process involves the movement of water from the soil through plant roots, up to the leaves, where it evaporates into the air. The energy required for this phase change from liquid to vapor is drawn from the surrounding environment, reducing both surface and air temperatures. Latent heat flux represents the amount of energy used in this transformation. Unlike sensible heat flux, which directly increases air temperature, latent heat flux removes heat and lowers the ambient temperature (Allen et al., 1994).

Meili et al. (2021) found that well-watered urban trees can decrease local air temperature by up to 5.8°C during summer due to evapotranspiration. However, the effectiveness of this cooling process depends on environmental factors such as water availability and vapor pressure deficit (VPD). When VPD is high, trees may limit transpiration by partially closing their stomata, reducing cooling efficiency during peak heat hours. Furthermore, Wu et al. (2024) found that evapotranspiration plays a secondary role compared to solar radiation interception.

Additionally, trees influence urban humidity through evapotranspiration, which increases moisture levels in the air. This process has a complex impact on thermal comfort: while higher humidity can reduce the efficiency of sweat evaporation - potentially making the environment feel warmer - evapotranspiration also helps lower air temperatures. However, Yang et al. (2024) argue that the net cooling effect ultimately improves thermal comfort. Further supporting this, Wang et al. (2015) found that trees improve thermal comfort in Assen, the Netherlands, based

on the Predicted Mean Vote model. Their study showed that on hot days, thermal perception decreased by 16% on clear days and by 11% on cloudy days due to the cooling benefits provided by street trees.

2.3. Review of SV data applications in temperature modelling

Multiple studies have shifted to focus from these physiological properties to the observed impact on street temperatures at a city scale. Three commonly used methods for studying this impact are in-situ measurements, satellite imagery, and SV data. SV data offers a scalable approach and provides a more accurate representation of the 3D urban structure by capturing multiple vertical layers that are often not visible in satellite images (Lu et al., 2023; Zhang et al., 2019). This research primarily focused on SV data due to these advantages. The section begins by clarifying commonly used temperature metrics, followed by an overview of related research using SV data for urban heat modelling. Finally, existing research gaps are summarized based on the evaluated studies.

2.3.1. Overview of temperature metrics

Three key temperature metrics are commonly referenced in literature. The first is air temperature, which represents the temperature of the air at a specific location. It is typically measured at a height of 1.25m or 2m above ground using thermometers at meteorological stations.

The second is LST, which refers to the temperature of the Earth's surface. LST is influenced by various factors, including solar radiation, surface material properties, and architectural morphology - such as building density, height, sky VF (SVF), and road surface area. Additionally, topographic elements like slope, aspect, and underlying surface conditions influence LST (J. Guo et al., 2020; Liu et al., 2023). LST is primarily measured using satellite-based sensors equipped with thermal imaging bands (Bhandari et al., 2022; Georgiou & Akçit, 2017).

The third measure is mean radiant temperature (MRT), which represents the uniform temperature of an imaginary environment where radiant heat exchange between the human body and its surroundings matches that of the actual, non-uniform environment (Enescu, 2017; Li, 2016). Closely linked to thermal comfort, MRT is measured using black globe thermometers, infrared thermometers, or calculated based on surface temperatures and view factors (H. Guo et al., 2020; Hua Ge, 2013).

2.3.2. Street view data for urban heat modelling

Lu et al. (2023) reviewed the use of SV data in urban environment research, with a focus on greenery. They identified the Green View Index (GVI) as the most used metric for quantifying greenery and noted that the Sky View Factor (SVF) - measuring sky exposure - was also frequently calculated. Their review showed that most studies have concentrated on the cultural services of street greenery, while other ecosystem services remained underexplored. Although four studies addressed reduced SVFs due to tree shading, only two explicitly examined temperature regulation (Kim et al., 2022; Li et al. 2022).

Li and Ratti (2018) investigated the spatial distribution of tree shade in Boston, USA, using GSV data and building datasets. They calculated the SVF by extracting sky pixels from SV

images using an algorithm based on brightness and hue differences. Next, they reprojected the extracted masks into Fisheye images to derive SVF values. To isolate tree-induced shading, they modelled the SVF resulting solely from buildings and subtracted it from the total SVF, resulting from both trees and buildings. Their findings indicated that trees accounted for 24.61% of all shade within the city, contributing significantly to thermal comfort. Additionally, they identified spatial disparities, with tree shade coverage being lower in commercial and industrial zones than in residential and recreational areas.

Kim et al. (2022) estimated MRT at a city scale using GSV imagery for Seoul, South Korea. They used a ResNet-18 convolutional neural network to apply semantic segmentation on panoramic images. They extracted sky, tree, road, sidewalk and build objects, and obtained Intersection over Union (IoU) values of 97,4%, 78,6%, 75,4%, 66,8% and 67,42% respectively. Further, the images were reprojected to a Fisheye perspective, and VFs were calculated for the extracted features. Each VF was assigned albedo, absorption, and heating coefficients to model the physical properties. Sun-path calculations were performed for each location to estimate incoming shortwave and longwave radiation. By integrating these factors, MRT was estimated and compared to LST derived from Landsat 8. Spatial analysis of MRT patterns revealed high values in open spaces and compact low-rise building areas. Discrepancies between LST and MRT were found in locations with low-growing vegetation, where LST was lower than MRT, highlighting the mitigating effect of shrubs and grasses on LST. Finally, the study also signified the effect of both high SVF and west-east street orientation resulting in high MRT values.

Wei et al. (2024) explored the relationship between greenery and LST across different urban morphology types (UMTs) in Austin, Texas, using GSV-data. They used a deep learning network for semantic segmentation, achieving a test accuracy of 66.64%, with no other evaluation metrics provided. Greenery levels were quantified using GVI, and LST was derived from Landsat 8 data using the radiative transfer equation. The LST dataset, initially at a 100-meter resolution, was resampled to 30 meters. Their results revealed a significant negative correlation between GVI and LST, with an LST reduction of approximately 0.77°C per 0.1 increase in GVI. However, this relationship varied across UMTs.

Zhang et al. (2019) investigated the relationship between 3D urban form and LST in Phoenix, Arizona, using GSV imagery. They used both spherical (GSV-based Fisheye projection) and planar (satellite-based) land cover fractions to characterize urban environments. For spherical images, semantic segmentation was applied using a fine-tuned fully convolutional network. The model was fine-tuned to six surface classes (sky, tree, buildings, impervious surfaces, pervious surfaces, and non-permanent objects) and the classification accuracy was above 95%. However, no other metrics, were provided. LST data was retrieved from ASTER. Using GWR at the block group level, they found that spherical land cover fractions provided more accurate LST estimates than planar fractions, as they better captured the effects of tree canopies and vertical structures on shading and radiation exposure. Their results indicated that higher spherical building fractions were generally associated with increased LST during both day and night - likely due to denser and taller buildings trapping longwave radiation. Additionally, a 1% increase in tree canopy cover was associated with a reduction in daytime LST ranging from 0.11°C to 0.53°C using GWR.

Li et al. (2022) further examined the effects of street vegetation on LST in Beijing, China, using a PSPNet semantic segmentation network to distinguish between grass, shrubs, and trees in

GSV data. They obtained IoU values (See Appendix B for more information on evaluation metrics) for trees, shrubs, and grasses of 89.6%, 73.3% and 84.1%, respectively. LST data was derived from Landsat 8. To analyse spatial heterogeneity, they conducted a hot spot analysis of LST and greenery distribution before applying GWR to assess the cooling effects of different vegetation types. Their findings indicated that vegetation types influenced cooling effects differently across various urban areas, highlighting the spatial variability in vegetation-LST interactions. Furthermore, they noted the importance of accounting for other factors of influence.

2.3.3. Insights from related work and remaining research gaps

This overview of related research demonstrates that GSV data can be a powerful tool for studying the urban environment. Semantic segmentation has enabled the processing of large volumes of SV images across entire cities to extract VFs. However, to date, few studies have focused on the ecosystem services of street greenery extracted from SV data. Furthermore, limited distinctions have been made between different types of greenery, with GVI remaining the most calculated VF (Lu et al., 2023).

Within the context of urban heat modelling, only one study was found that employed a custom, fine-tuned model (Zhang et al., 2019). Other studies relied on pretrained models without further customization, resulting in limited performance or the inclusion of only a narrow range of relevant semantic classes (Kim et al., 2022; Li et al., 2022; Wei et al., 2024). It has also been shown that the extracted VFs from semantic segmentation models can be used to analyse relationships with shadowing, MRT, or LST using various statistical methods.

Regarding the relation between SV extracted features and LST specifically, existing studies have either grouped greenery into a single category or neglected other urban characteristics beyond greenery (Li et al., 2022; Wei et al., 2024). Additionally, current research has primarily relied on correlation analysis or GWR to quantify the relationship with LST (Li et al., 2022; Wei et al., 2024; Zhang et al., 2019). While GWR can account for spatial non-stationarity, it cannot capture non-linear or non-monotonic relationships. Moreover, it is sensitive to outliers, and its interpretability is not always straightforward (Subedi et al., 2018).

2.4. Semantic segmentation

Semantic segmentation through deep learning has been the most used method for extracting information from SV images. It offers both a scalable and accurate method, allowing for city wide applications. The process involves assigning a specific class label to each pixel in an image. Deep learning methods are particularly effective for this task due to their capacity to learn and use contextual information (Lu et al., 2023; Xiao et al., 2024). Semantic segmentation models consist of two main components: the model architecture and the dataset used for training, along with its corresponding classification scheme.

Within this study, Mask2Former was used to segment GSV images and extract urban features. Mask2Former is a universal architecture designed for three different segmentation tasks: panoptic, instance and semantic segmentation. Here, a brief overview of the architecture is provided to showcase the models working.

Mask2Former is an extension of the MaskFormer architecture, for which the model’s architecture is presented in Figure 1 (Cheng et al., 2022). With MaskFormer, a backbone network extracts the initial image features. These features are then upsampled by a pixel decoder module to obtain per-pixel embeddings (Figure 1, lower left). In parallel, the Transformer decoder processes the same extracted features to generate per-segment embeddings using a mechanism called cross-attention, which assigns different importance weights to different parts of the image. (Figure 1, upper left). These per-segment embeddings capture global contextual information, summarizing the overall shape and appearance of potential segments as prototype representations. Next, these prototypes are used to derive class labels and mask embeddings - each prototype segment is assigned a class label along with its corresponding mask embedding. Binary masks are produced by computing the dot product between the per-pixel embeddings and the mask embeddings, followed by a sigmoid activation function. This operation results in masks that can potentially overlap for each object instance. Finally, to achieve semantic segmentation, the generated objects are integrated with their respective class predictions, resulting in the final segmented image (HannaMergui, 2023a).

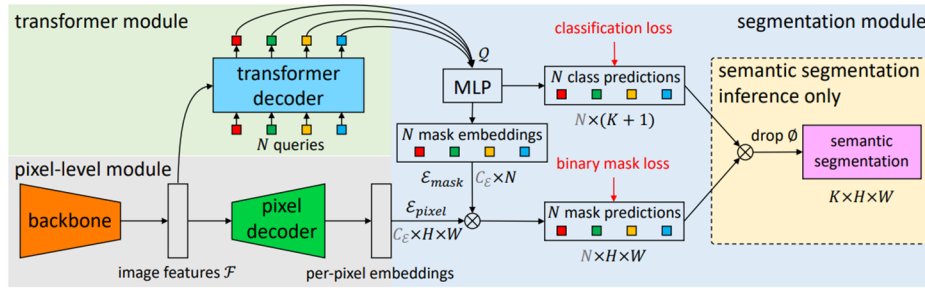


Figure 1: MaskFormer Architecture. Note. From “Per-Pixel Classification is Not All You Need for Semantic Segmentation,” by B. Cheng et al., 2021, arXiv (Cornell University).

The architecture of Mask2Former is different from its predecessor in two ways. First, it implements a decoder with masked attention instead of cross-attention. With masked attention, only a portion of the input sequence is focused on, leading to faster convergence and improved performance. The second improvement involves using multiscale feature representation to capture both fine-grain details and contextual information. These multi-scale features are processed by feeding them to one Transformer decoder layer at a time, ensuring that each layer specializes in processing features at a specific scale (Figure 2) (Cheng et al., 2022; HannaMergui, 2023b).

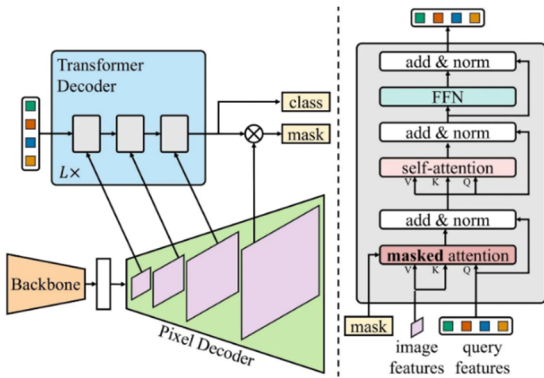


Figure 2: Mask2Former architecture. Note. From “Masked-attention mask transformer for universal image segmentation,” by B. Cheng et al., 2022, arXiv (Cornell University).

2.5. Machine learning

For modelling the relationship between urban features and LST this research relied on different machine learning techniques. Random forest, and its geographical extension, GWRF, were used to analyse the contributions of urban features on LST and respective spatial patterns.

2.5.1. Random forest

Random forest is an ensemble learning method that combines the predictions of multiple decision trees to produce a final prediction. Each decision tree is trained on a bootstrap sample of the data (a randomly selected subset of the data with replacement) and constructed by recursively splitting this data into smaller subsets. At each split, a random subset of features is considered, and the optimal split is determined by the feature that best partitions the data into more homogeneous subsets, as quantified by an impurity metric. Splitting continues until a certain threshold of impurity decrease is reached, or until further splitting is restricted by other pre-set hyperparameters (Breiman, 2001; Salman et al., 2024).

Random forest offers several advantages. It can be used for both classification and regression problems, capable of handling large datasets with many variables. Further, the algorithm can model complex, non-linear relationships among variables. By using bootstrap sampling and random feature selection at each node, the correlation between individual trees is reduced, improving the overall combined predictive performance. Additionally, random forest provides estimates of feature importances, improving the interpretability of the model predictions (Breiman, 2001; Salman et al., 2024).

2.5.2. Geographically Weighted Random Forest

GWRF integrates a globally trained random forest with locally trained models to predict each instance in the dataset. The global model captures general trends by using the full dataset, while the local models capture spatial heterogeneity in the dataset. By combining global and local predictions, the approach can improve model performance and interpretation and provide more reliable estimates of feature importances by incorporating latent non-stationary effects (Georganos et al., 2019).

In GWRF, the locally trained models are created by iterating over every instance within the dataset and creating a local random forest using spatially close features. The number of features included in the local random forest is defined by the bandwidth, which can either be fixed or adaptive. For an adaptive bandwidth, the nearest number of instances is included in the model, while for a fixed bandwidth all features within a certain distance are included. Furthermore, weights can be assigned to all features within a specified bandwidth using a weighting kernel and the distances of the included instances to the central point in the local model. Examples of kernels include bi-square, gaussian and exponential. Using these weights, more emphasis can be put on spatially nearer features (Georganos et al., 2019; Georganos & Kalogirou 2022). The bandwidth for GWRF can be optimized using Out-Of-Bag (OOB) error minimization, a method introduced by Georganos & Kalogirou (2022). This approach involves generating OOB predictions for the target point of each local model and comparing them to the actual observations. By repeating this process across a range of bandwidth values, it is possible to obtain an unbiased estimate of the model performance for each bandwidth and thus select the most optimal one.

To integrate the results of the global and local models, prediction weighting is applied in GWRF following Equation 1 (Georganos et al., 2019):

$$y_i = (1 - \alpha)y_{g,i} + \alpha y_{l,i} \quad (1)$$

Where:

- y_i is the final prediction for location i ,
- $y_{g,i}$ is the prediction from the global model at location i ,
- $y_{l,i}$ is the prediction from the local model at location i ,
- α is the local weighting parameter.

The parameter α determines the relative contribution of the global and local models and varies within the interval [0,1]. Two main approaches have been described in the literature for selecting an appropriate α . The first involves varying α to minimize the overall prediction error of the combined model (D. Wu et al., 2023; Georganos et al., 2019). The second is a theoretically informed approach based on the global Moran's I statistic computed for a specific bandwidth. Since Moran's I quantifies the degree of spatial autocorrelation, it can be used to guide the weighting. A higher Moran's I indicates stronger spatial autocorrelation, suggesting that greater weight should be assigned to the local model. Conversely, a lower Moran's I implies weaker spatial autocorrelation, suggesting that greater weight should be assigned to the global model (Sun et al., 2024).

2.5.3. Interpretability of model predictions

Within the context of this research, the focus was on understanding how model predictions were generated to understand the relationships between different urban characteristics and LST. This required a method capable of transparently analysing the behaviour of the random forest and GWRF models. SHAP values provide local explanations for machine learning predictions by attributing an additive importance score to each independent variable. This results in a unique and consistent measure of variable contribution. Furthermore, due to their additive nature, SHAP values share the same unit as the dependent variable, allowing direct interpretation of each independent variable's influence (Lundberg et al., 2017).

3. Method and Materials

The overall flow of the methodology is depicted in Figure 3. LST data was collected using a tool for upscaling thermal satellite imagery to a higher resolution, with Landsat 8 imagery as input. GSV data was retrieved by sampling the street network and followed by retrieving the GSV data using the point coordinates. Two methods were applied for retrieving GSV data: Multi-View Collection (MVC) and Panoramic Image collection (PIC). MVC included capturing the hemisphere by collecting individual images and PIC capturing the hemisphere by retrieving 1 panoramic image. Additionally, the area covered by water within a radius of 800m and street orientation were retrieved for every sampled location.

To extract features of interest from GSV images, a fine-tuned semantic segmentation model was used. Fine-tuning involved manually annotating a subset of MVC images and training the model on this data. The trained model was then applied to segment both MVC and PIC imagery,

enabling the calculation of VFs to quantify the relative presence of trees, shrubs, grass, buildings, and sky exposure at each location (research question 1).

To compare the three commonly used VF quantification methods, Fisheye imagery was generated by reprojecting PIC-based VFs. So, this resulted in VF estimates for three image types: MVC, PIC, and Fisheye. These were then compared in terms of magnitude, spatial distribution, and their influence on model behaviour and performance for predicting LST (research question 2).

To model the relationship between urban characteristics and LST, GWRF was used with the MVC-derived VFs, street orientation, and water area within an 800 m as inputs. SHAP analysis was applied to interpret the model and assess variable importance (research question 3).

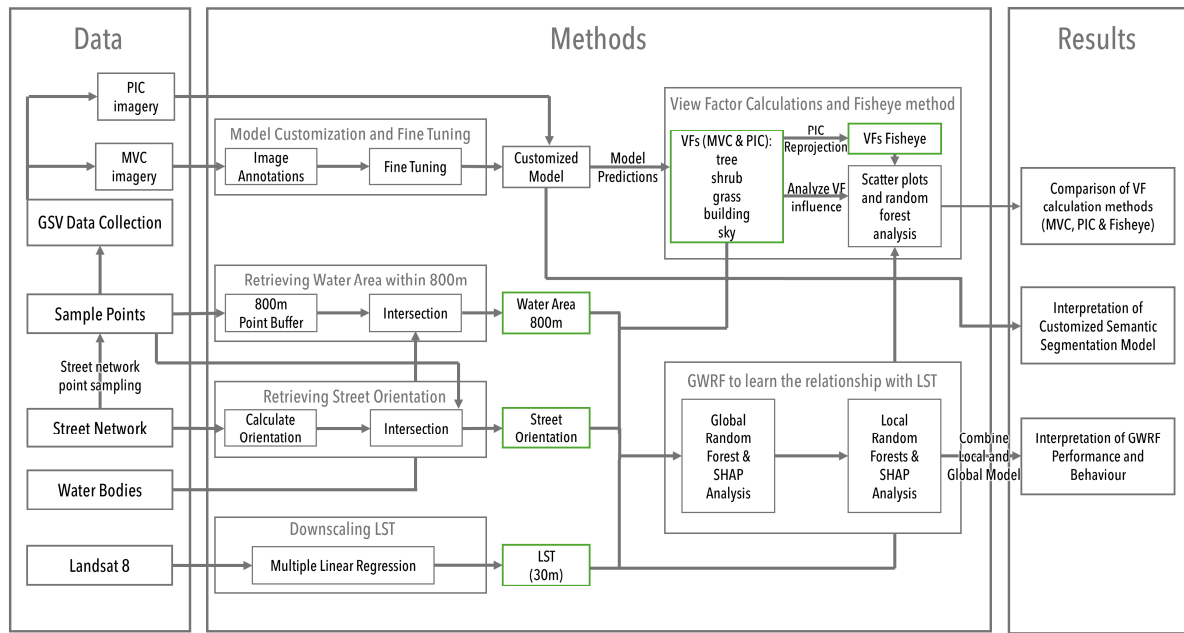


Figure 3: Flowchart of the research.

3.1. Study area and data collection

3.1.1. Study area

Zurich was selected as a case study for this research. Zurich is Switzerland's largest city with 448,664 inhabitants. It is the capital of the Canton of Zurich and is located in the north central part of the country, at the north-western tip of Lake Zurich. The municipality has a total area of 91.88 km², of which 4.1 km² is occupied by the lake. The city is divided in two by the Limmat River, which flows north from the lake. Zurich's climate is classified as Cfb according to the Köppen climate classification, indicating a temperate oceanic climate. This classification is characterised by the coldest month averaging above 0°C, all months averaging below 22°C, and at least four months with average temperatures above 10°C. The Cfb climate zone is the same as that of several other major European cities, including Amsterdam, Berlin, Brussels and Paris. This makes Zurich a well-suited case study regarding the representativeness of temperature dynamics (Wikipedia contributors, 2025; Millison, 2019) (Figure 4).

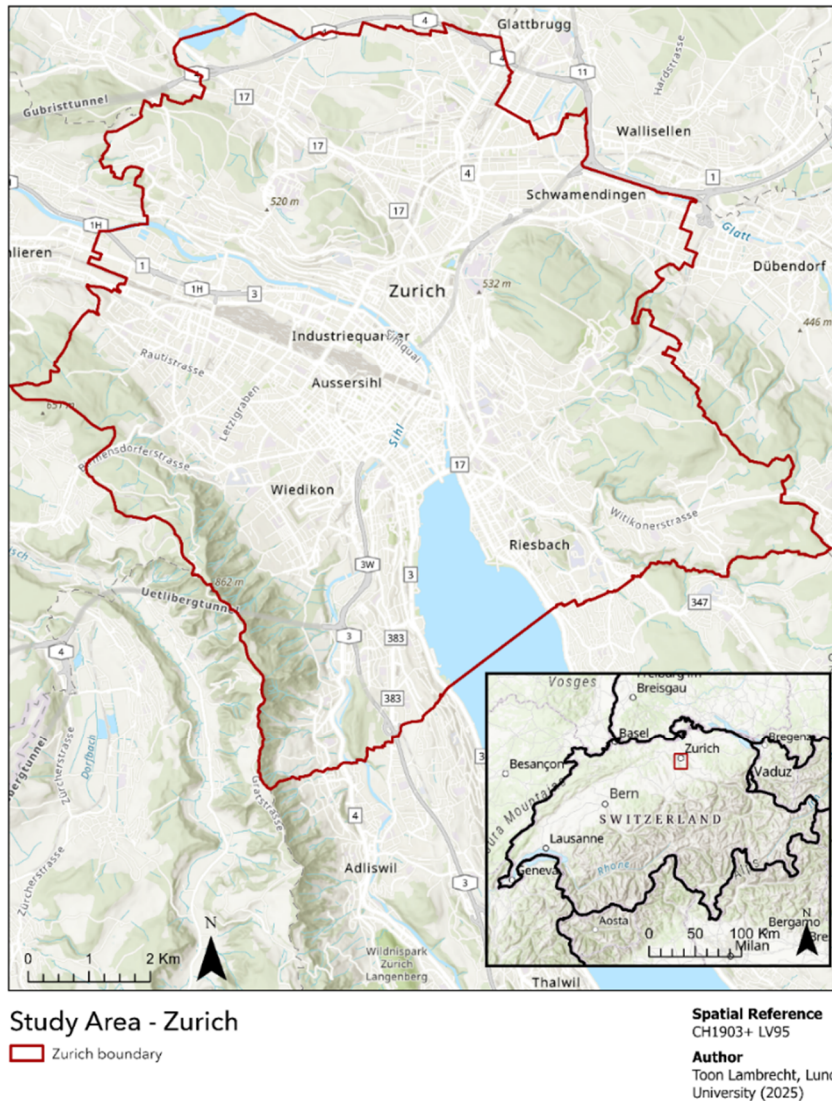


Figure 4: Overview the study area – Zurich.

3.1.2. Street view data collection

The street network of Zurich was retrieved from the Topographic Landscape Model of Switzerland (*The Topographic Landscape Model (TLM)*, n.d.). A subset of the road database in TLM was extracted by filtering out smaller roads that were inaccessible to cars. Along this simplified road network, points were sampled at 50-meter intervals. 50 meters was chosen to have a sufficiently dense sample network, while also preventing large overlap between images. This resulted in a total of 16,285 sample points.

GSV images were collected using the Street View Static API, which operates via HTTPS requests and offered two main functionalities. First, metadata could be retrieved by sending a request URL specifying a location in WGS84 coordinates. The response was a JSON file that indicates the availability of GSV images for the specified location. If images were available, the response included a panoramic ID along with the date of recording (Google Maps Platform Documentation, n.d.). While some small clusters of locations lacked image coverage, these gaps were dispersed throughout the road network. The ratio of unavailable to available images was

0.85, indicating that data was available for most locations (13,801). Using the available panoramic IDs, images were retrieved for all locations for the MVC and PIC methods. The spatial distribution of availability is shown in Appendix A.

3.1.2.1. Multi-View Collection of GSV imagery

For MVC, 12 images were captured per location by varying two parameters: compass heading (horizontal angle: 0° , 90° , 180° , and 270°) and pitch (vertical angle: -45° , 0° , and 45°). The field of view (FOV), which defines the horizontal coverage, was set to 120° (Google Maps Platform Documentation, n.d.). Like metadata retrieval, an HTTPS request was sent to the server to obtain an image for every location and parameter pair. For example, the URL for location 11665, with a heading of 0° , pitch of 0° , and an FOV of 120° , was composed like: https://maps.googleapis.com/maps/api/streetview?size=600x400&pano=Ne74rTk2IUkGpWMKiF_YNw&heading=0&pitch=0&fov=120&radius=50&source=outdoor&key={personal_api_key}. The result of this request is shown in Figure 5.



Figure 5: API response to request for location with FID 11665, heading 0, pitch 0 and FOV 120.

3.1.2.2. Panoramic Image Collection of GSV imagery

For PIC, the same panoramic IDs were used to ensure that the retrieved images were captured at the same moment as those in MVC. However, instead of retrieving individual images, the Python module developed by Cplusx (2019) was used to obtain full panoramic images, covering approximately the entire visible hemisphere. This tool stitched together 26×13 tiles per location into a single panoramic image. The result for location 11665 is shown in Figure 6.



Figure 6: Panoramic image for location with FID 11665.

3.1.3. Temperature data collection

Before retrieving LST data, the temporal dimension of the GSV data was investigated. As mentioned in section 3.1.2, the GSV images were retrieved using a panoramic ID. Since each panoramic ID was linked to a timestamp in the ‘year-month’ format, the temporality of the dataset was analysed by first reclassifying each timestamp into a season and then plotting the corresponding distribution. Most of the data was concentrated in the summer and fall of 2024, with all fall 2024 images specifically captured in September. Another peak was observed in the summer and fall of 2022 (Figure 7).

The fact that GSV images were captured over different years and seasons introduced certain challenges. Since plant phenology varies with the seasons, the detection of greenery may be biased. Images taken when trees had full canopies were likely to display a higher proportion of greenery pixels, resulting in a positive bias. Additionally, yearly variations posed a problem: comparing greenery detected in images from 2022 with LST data from 2024 could overestimate the influence of greenery on a location, as plants may have grown during the two-year gap, thereby enhancing their mitigating effect. However, given that urban greenery is typically pruned regularly, this effect was assumed to be less pronounced. Consequently, LST imagery was collected for the summer of 2024.

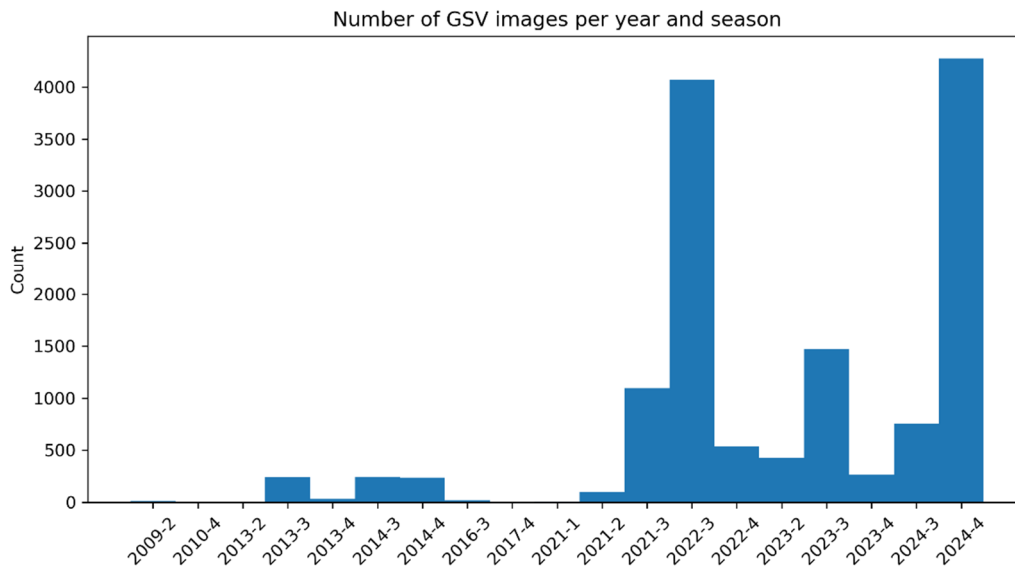


Figure 7: Temporal dimension of the GSV data (1: Winter; 2: Spring; 3: Summer; 4: Fall).

Temperature data was collected using the tool developed by Onáčillová et al. (2022). The tool used a downscaling approach to calculate LST at a 30m resolution by integrating LST data from Landsat 8 (100m resolution) with the NDVI, Normalized Difference Water Index (NDWI), and Normalized Difference Building Index (NDBI). These parameters were combined in a multiple linear regression model and the tool was implemented in Google Earth Engine (GEE), enabling users to easily retrieve LST data for a specified period and region at a 30m resolution. In total, seven images with less than 15% cloud coverage were collected. Additionally, the images were manually inspected to identify cloud-covered areas and none of the images contained clouds over the city itself. Finally, the images were averaged on a pixel-by-pixel basis, to get a better

estimate of overall summer temperatures (Figure 8). From the resulting averaged raster layer, pixel values were extracted at the sample point locations for which GSV imagery was collected.

On the collected LST data, urbanized areas were distinctly visible, showing clear temperature contrasts compared to surrounding non-urbanized regions. Moreover, variability was observed within the urban landscape. Localized patches of lower LST aligned with larger parks and water bodies. For waterbodies, both the lake and the river flowing through the city were clearly visible, displaying their cooling effect. Inter-street temperature variability was more subtle but also observed (Figure 8).

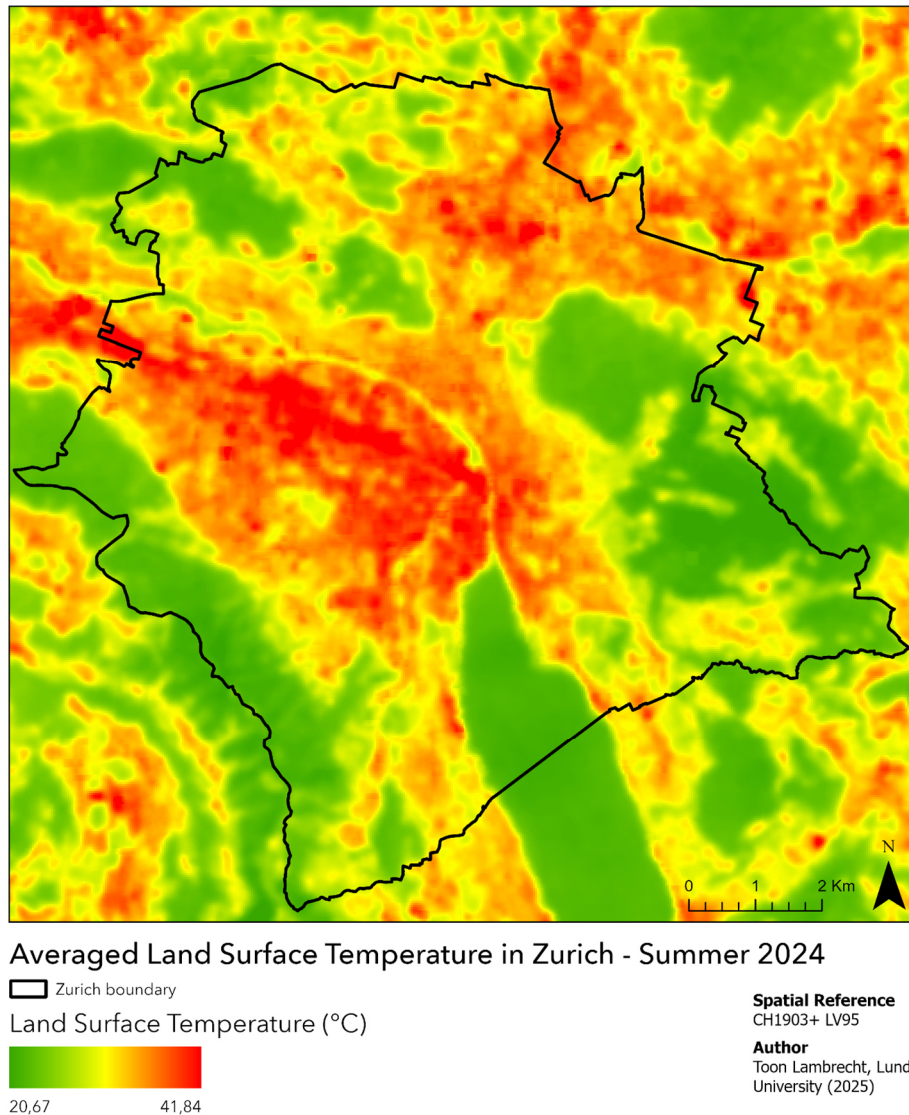


Figure 8: *Averaged Land Surface Temperature in Zurich for Summer 2024.*

3.1.4. Additional drivers of LST

In addition to buildings, trees, shrubs, grass and sky-exposure - that were extracted through semantic segmentation (see section 3.2) - two other factors were collected: street orientation and proximity to water. Kim et al. (2022) emphasized the importance of street orientation for estimating MRT, while Wu and Zhang (2019) found that water bodies can influence LST up to 800 m. Since both variables have been shown relevant, and Zurich contains significant water

bodies, including a lake and several rivers (see Figure 4), the variables were derived from existing datasets.

Street orientation was determined by segmenting the road network into smaller parts, so that each pair of connected vertices formed a street segment. The compass direction of each segment was then calculated using the coordinates of the vertices. These line values were linked to the sampling points through spatial intersection and finally, the compass directions were reclassified into East–West (EW) and North–South (NS) categories based on the classification scheme shown in Figure 9.

For water bodies, the focus was shifted from simply distance to water, to quantifying the surface area covered by water within a certain distance. This was done by extracting lakes and rivers from the Swiss TLM (TLM, n.d.). Line features representing watercourses were manually converted into polygons to fully cover the river areas, using satellite imagery as a reference. For each sampling point, an 800 m buffer was created and intersected with the water bodies layer. The area of intersection was then divided by the total buffer area to calculate the proportion of water surface within 800m.

Street Angle Reclassification with Breaks

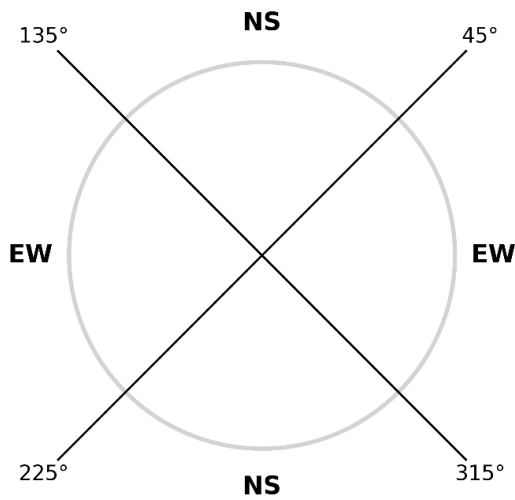


Figure 9: *Reclassification scheme for street angles. All angles between 45° - 135° and 225° - 315° were reclassified to NS and all angles between 45° - 315° and 135° - 225° were reclassified to EW.*

3.2. Semantic segmentation

In the previous section, the collection of GSV imagery, LST data, and the two additional influencing factors was described. This section outlines the process of performing semantic segmentation to extract the five features of interest from the collected GSV images. It details how a deep learning model was selected, modified, fine-tuned, evaluated, and subsequently used for inference. This process was done to address research question 1, which examined whether the urban environment can be accurately mapped focusing on multiple features.

3.2.1. Model selection

Different semantic segmentation architectures were explored for feature extraction. Mask2Former was chosen based on its strong performance, unified architecture, availability of pretrained models, and the existence of established frameworks for model use and training. Facebook provided a diverse range of pretrained Mask2Former models in its model zoo, trained on various benchmark datasets and using backbone models of varying complexity (Cheng et al., 2022; *Mask2Former*, n.d.). In the context of computer vision tasks, benchmark datasets are standardised, publicly available datasets that are commonly used to evaluate and compare machine learning models' performance on specific tasks.

A key consideration was selecting a model based on the representativeness of semantic classes in these benchmark datasets. Three commonly used benchmark datasets were Cityscapes, COCO, and ADE20K. Cityscapes included 30 classes, all relevant to urban environments, however, it generalised vegetation into a single class. COCO-Panoptic comprised 133 classes, including two vegetation-related categories: *tree-merged* and *grass-merged*. ADE20K, with 150 classes, offered more granularity in vegetation classification, distinguishing grass, trees, and shrubs as separate categories. However, the high number of urban-specific classes in ADE20K introduced additional complexity and potential confusion in object classification.

Given that the pretrained COCO-Panoptic model classes aligned more closely with the relevant categories in this study, the *mask2former-swin-large-coco-panoptic* model was selected from the model zoo. However, since it lacked a distinct shrub class - considered important for the analysis - the COCO-Panoptic classification scheme was modified to introduce *shrub-merged* as a third vegetation class. The impact of this modification on model performance was then evaluated.

3.2.2. Model classes of interest

The target classes to be extracted from the GSV imagery were trees, shrubs, grass, buildings, and sky. The selected model included multiple classes that corresponded to these five generalized classes. Any classes that did not semantically align with these categories were classified as background. An overview of how the original COCO-panoptic classes were mapped to the five generalized classes is provided in Table 1. Image annotations were made using the original COCO-panoptic classes (section 3.2.3), while for model evaluation (section 3.2.6) the analysis focused exclusively on the generalized categories, since these were relevant to the research question.

As outlined in section 3.2.1, the model did not include a shrub class within the classification scheme. *Shrub-merged* was therefore introduced by creating a new category, following these operational definitions for the vegetation categories:

- *grass-merged*: Includes low-growing, non-woody vegetation such as lawns, meadows, and other ground-covering grasses. It consists of sparse to moderately dense vegetation that does not exhibit significant vertical growth or woody stems.
- *shrub-merged*: Everything denser than grass, including small bushes, ground cover, and hedgerows. Shrubs are typically woody plants with multiple stems and a height generally lower than trees, forming a mid-layer in vegetation.

- *tree-merged*: Includes all tall, woody vegetation with a distinct trunk or multiple large stems extending upward to form a canopy. This category covers individual trees as well as tree groups, regardless of seasonal foliage presence.

Table 1: Overview of how COCO-panoptic classes were reclassified into the 5 generalised categories.

COCO-panoptic classes	Generalised classes
tree-merged	<i>tree-merged</i>
manually introduced	<i>shrub-merged</i>
grass-merged	<i>grass-merged</i>
building-other-merged, house	<i>building-merged</i>
sky-other-merged	<i>sky-merged</i>
all other classes	<i>background-merged</i>

3.2.3. Image Annotation

Fine-tuning a segmentation model required the creation of training data, which was achieved by manually classifying pixels using the COCO-Panoptic classification scheme. A subset of 400 GSV images was selected from the acquired MVC dataset and uploaded to the CVAT.ai server, an open-source annotation tool (CVAT.ai, n.d.). Using CVAT.ai, 358 of these images were manually annotated by delineating objects with pixel masks. Two annotation datasets were created, one maintaining the classification scheme of COCO-panoptic, and one with the modified classification scheme containing *shrub-merged*.

In some cases, objects in the background were difficult to identify due to image resolution or object size. Additionally, certain objects did not fit into any category within the COCO-Panoptic classification scheme. To ensure high-quality training data, these pixels were left unannotated in both cases. In total, unannotated pixels accounted for only 0.34% of the total dataset, minimizing their impact on model training.

3.2.4. Data Preparation

The transformers library was used for fine-tuning the model. The library and implemented tuning process used data in a specific format, which necessitated different preprocessing steps to correctly format the data, starting from the annotations made in cvat.ai.

The image annotations within the exported document from cvat.ai were transformed into segmentation masks using the Datumaro library (Datumaro, 2024). The resulting segmentation masks followed a custom formatted RGB colour scheme, with one distinct colour per annotation class. Unidentifiable pixels, or pixels that didn't belong to a specific class, didn't have any colour and were left black. Figure 10 displays the resulting segmentation masks for image location 8123. The annotated dataset - consisting of images and respective segmentation masks – was split into 60/20/20 training, validation, and testing subsets and organized into a dataset object (*The Dataset Object*, n.d.). This was done for both the annotated dataset with and without *shrub-merged*.

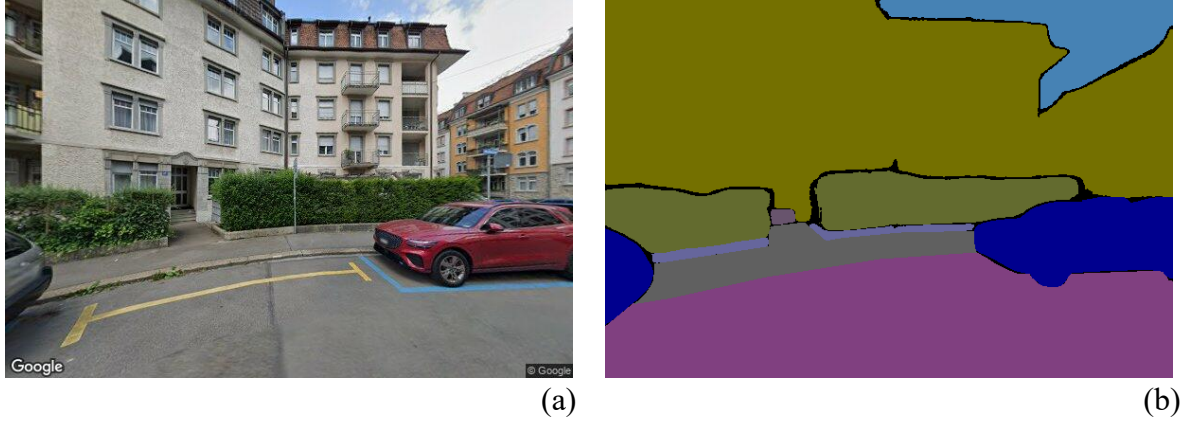


Figure 10: (a) represents the actual image and (b) the manually created segmentation map using CVAT.ai. The image corresponds to location 8123, with heading 270° and pitch 0°.

3.2.5. Model fine tuning

After the training data was created, the selected model was fine-tuned to customize it to the city of Zurich. It was once fine-tuned using the dataset with the original classification scheme (Model 1) and once fine-tuned using the dataset with the classification scheme including *shrub-merged* (Model 2). The fine-tuning process consisted of three main stages: creating a PyTorch dataset with custom data augmentations, hyperparameter tuning, and model fine-tuning.

The first step in fine-tuning the model was to create a PyTorch dataset with custom data augmentations. These augmentations are techniques that artificially increase the diversity of the training data by applying transformations to it. This improves generalisation and mitigates overfitting. Defining the dataset allowed images and their corresponding segmentation labels to be loaded while applying the custom augmentations. The augmentation techniques applied included resizing, horizontal flipping, image rotation and normalisation.

Hyperparameter tuning was the second step in the process of fine-tuning the model. Hyperparameters are configurational variables that are defined before the learning process starts. These variables control how the model learns and are critical to optimising model performance. Hyperparameter tuning is the process of identifying the optimal set of hyperparameters to achieve the best possible validation performance during training. This involves evaluating multiple combinations of hyperparameters and selecting the set that yields the highest performance (Owen, 2022).

Table 2 presents the hyperparameters sets used for evaluation, along with their respective intervals or discrete options. The evaluation metric during tuning was *loss*, calculated using a weighted function that combines sigmoid cross-entropy and dice loss. To achieve optimal model performance, this metric had to be minimised for both the training and validation datasets. The number of epochs was set to three during hyperparameter tuning, meaning the optimal parameters were those that minimized validation loss after three epochs. The total number of trials was set to 60, with an additional setting to ensure that unpromising trials were deprecated early, to minimise the computational demand. During every trial, a hyperparameter combination was sampled, using log-uniform sampling for continuous variables, and uniform sampling for discrete values.

The learning rate (y) was defined using an exponential function given by Equation 2.

$$y = a(1 + r)^t \quad (2)$$

Where:

- a represents the initial learning rate,
- r is the decay factor, and
- t denotes the epoch.

Using an exponential function ensures that the learning rate is initially high, allowing the model to learn quickly, but gradually decreases over time to prevent overshooting and improve convergence.

Table 2: *Hyperparameter tuning search space*

Hyperparameter	Type	Values
Initial Learning rate	Continuous	$[9 * 10^{-6}; 1 * 10^{-3}]$
Decay factor	Continuous	[0.3; 0.9]
Batch Size	Discrete	[2; 4; 6]
Optimizer	Discrete	[Adam; SGD; RMSprop; AdamW]

Finally, after hyperparameter tuning, the model was fine-tuned using the optimal identified hyperparameters. Throughout training, both training loss and validation loss were monitored. Training was terminated if the training and validation losses began to diverge, or if the validation loss started to increase, indicating overfitting.

3.2.6. Model evaluation

After fine-tuning the model, evaluation was done for both Model 1 and Model 2. Model evaluation was performed by comparing the predicted values with the annotated ground truth values in the test splits. Unclassified pixels in the test labels were filtered out to ensure that only pixels with known values were included in the evaluation. Further, evaluation was done for both Model 1 and Model 2, and evaluation metrics were calculated for all generalised categories (Table 1). Due to class imbalances in the test dataset, three metrics were primarily focused on for model evaluation: IoU, precision (P), and recall (R). (A comprehensive overview of these metrics is provided in Appendix B.) In addition to these metrics, row-normalized confusion matrices were computed for Model 1 and Model 2. These matrices provided a breakdown of misclassification patterns, offering insights into which categories the model tended to confuse. Row normalisation was applied to ensure that the values in each row represented the proportion of actual instances of a given class that were predicted as each possible class. This made it easier to interpret the model’s performance, as it emphasised the distribution of predictions for each true class, independent of class frequency.

3.2.7. Model inference

After evaluating the model, all images were predicted to extract urban features at each location. As outlined in Section 3.1.2, two different methods were used for collecting GSV images (MVC & PIC). Model 2 - including *shrub-merged* - was applied to predict the images from both methods. However, despite using the same model, there was a slight variation in the prediction process.

For images collected using MVC, each image was individually processed by the model and stored. For images collected using PIC, the panoramic images were first split in half to better align with the input dimensions of the training data (MVC) used for fine-tuning the model. Each half was then processed separately. After prediction, the two halves were recombined before storing the final predictions.

3.3. View factor calculations

VFs were used as a tool to characterize the urban environment at specific locations and were calculated for the 5 generalized classes (Table 1, excluding *background-merged*) extracted during model inference. To calculate the VFs, three different methods were applied. While these methods have been used individually in the literature, no comparative studies have yet focused on the interrelationships between the calculated VFs or their influence on model behaviour and predictive accuracy (Li & Ratti, 2018; Wei et al., 2024; X. Li et al., 2015, Lu et al., 2023). To address this, the three different methods were summarised in this study and applied.

For the MVC approach, a single method was applied for VF calculation, following the procedure outlined by X. Li et al. (2015). For the PIC data approach, two different methods were employed. The first method directly derived VFs from the predictions of the PIC images, while the second method involved reprojecting the predictions into a Fisheye perspective prior to VF calculation.

The following sections first describe the methodology used for MVC, then the methodology applied to PIC, and finally, the reprojection process and subsequent VF calculation from the Fisheye images.

3.3.1. View factor calculations for MVC

For the MVC method, each location consisted of 12 images, corresponding to four compass headings (north, east, south, and west) and three pitch angles (upward, horizontal, and downward). The VF for each class was calculated as the proportion of the number of pixels classified into that class relative to the total number of pixels across all images, and is described by Equation 3 (X. Li et al., 2015):

$$VF_x = \frac{\sum_{i=1}^4 \sum_{j=1}^3 N_{x_{ij}}}{\sum_{i=1}^4 \sum_{j=1}^3 N_{total_{ij}}} \quad (3)$$

Where:

- VF_x represents the view factor for class x ,
- $N_{x_{ij}}$ is the number of pixels classified as x for heading i and pitch j ,

- $N_{total_{ij}}$ is the total number of pixels in the image for heading i and pitch j .

3.3.2. View factor calculations for PIC

For PIC, the same fundamental principle was followed as for MVC, however, since there was only one panoramic image, Equation 3 simplified to Equation 4 (B. Li et al., 2022; Wei et al., 2024):

$$VF_x = \frac{N_x}{N_{total}} \quad (4)$$

Where:

- VF_x represents the view factor for class x ,
- N_x is the number of pixels classified as x ,
- N_{total} is the total number of pixels in the image.

3.3.3. View factor calculation for Fisheye imagery

Lastly, VFs were calculated using Fisheye imagery. These images were generated by reprojecting panoramic images (PIC), which cover the entire hemisphere, onto a circular plane to create a fisheye perspective simulating a “look-up-from-the-ground” view. This projection type has been widely used in studies characterizing the urban environment, particularly for calculating the SVF. The use of fisheye projections in these studies was done to better approximate how the urban environment is perceived from ground level. Additionally, this method has been argued to provide a more accurate estimation of incoming solar radiation, as the images are directly oriented towards the sky (Gong et al., 2018; Kim et al., 2022; X. Li & Ratti, 2018; Lu et al., 2023).

To obtain the Fisheye VFs, the PIC images were first reprojected to Fisheye images using the approach of X. Li and Ratti (2018). This approach involved constructing a mathematical relationship between pixels on a Fisheye image (x_f, y_f) and pixels on a PIC image (x_p, y_p). This relationship was defined by Equations 5, 6 and 7:

$$x_p = \begin{cases} \left(\frac{\pi}{2} + \tan^{-1} \left[\frac{y_f - C_y}{x_f - C_x} \right] \right) * \frac{W_p}{2\pi}, & x_f < C_x \\ \left(\frac{3\pi}{2} + \tan^{-1} \left[\frac{y_f - C_y}{x_f - C_x} \right] \right) * \frac{W_p}{2\pi}, & x_f > C_x \end{cases} \quad (5)$$

$$y_p = \left(\frac{\sqrt{(x_f - C_x)^2 + (y_f - C_y)^2}}{r_0} \right) * H_p \quad (6)$$

$$r_0 = \frac{W_p}{2\pi} \quad (7)$$

Where:

- W_p and H_p represent the width and height of the PIC image,
- r_0 is the radius of the fisheye image,
- (C_x, C_y) are the coordinates of the center pixel in the fisheye image.

The result of reprojection is shown in Figure 11.



Figure 11: Reprojected fisheye image of location 11665. Note, this is for the same location as Figure 5 and Figure 6.

After reprojection, the VFs were calculated by dividing the fisheye image into concentric annuli of equal width to approximate the integral over the width of the Fisheye image. The VFs for each category x were then obtained by summing the contributions of all annular sections, using Equation 8 (Johnson & Watson, 1984; X. Li & Ratti, 2018):

$$VF_x = \frac{1}{2\pi} \sin\left(\frac{\pi}{2n}\right) \sum_{i=1}^n \sin\left[\frac{\pi(2i-1)}{2n}\right] \alpha_{i,x} \quad (8)$$

Where:

- VF_x represents the view factor for class x ,
- n is the number of rings,
- i is the index of the ring,
- $\alpha_{i,x}$ is the angular width of pixels for feature x , in the i th ring.

3.3.4. View factor comparison

The VFs were compared by plotting them for each class and the three methods against each other along the identity line. Additionally, the coefficient of determination (R^2), slope, and intercept were calculated to determine the best-fit relationship between the two methods (More information on R^2 is provided in Appendix B). The identity line (also called 1:1 line) represents the ideal scenario in which both methods produce identical VF values, meaning all data points would fall precisely along this line in the case of perfect correspondence.

To further compare the three methods, three random forest models were built using five-fold cross-validation. From these models, feature importances were extracted to assess whether the VF calculation method affected model behaviour. In addition, the Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 were computed for each model to compare the overall performance (See Appendix B for a comprehensive overview of these metrics). Street

orientation and water area were also included in the random forests to ensure proper comparison across models and to ensure that the same parameters were used as for the final GWRF model for learning the relationship with LST (section 3.4).

3.4. Exploring the relationships between urban features and LST

After collecting LST data and the influencing factors - including 5 VFs for *trees-merged*, *shrub-merged*, *grass-merged*, *building-merged*, and *sky-merged*, as well as street orientation and water area within an 800 m radius - GWRF was applied to model the relationship between the dependent and independent variables (lower right of flow chart in Figure 3).

3.4.1. Training the global model

To prevent overfitting and optimise model performance, the global model was trained using hyperparameter tuning with five-fold cross-validation. Grid search, a method that evaluates all possible combinations of specified hyperparameter values, was used within a constrained parameter space to minimise the risk of overfitting the training data. The details of the parameter space are presented in Table 3.

Table 3: Hyperparameter search grid for the global random forest.

Parameter	Number of estimators	Maximum depth	Minimum samples per split	Minimum samples per leaf	Maximum features
Range	[200, 300, 400]	[8, 10, 12, 14]	[7, 9, 11]	[7, 9, 11]	[sqrt, log2]

3.4.2. Training the local models

To prevent data leakage - the unintentional use of test data during model training- each local model was trained by excluding its corresponding target point. Given the continuous nature of temperature data, a minimum of 200 neighbours was set during OOB bandwidth optimisation using an adaptive bandwidth approach with a bi-square kernel. This threshold was chosen to mitigate the risk of overfitting. Specifically, 200 neighbours typically corresponded to bandwidth distances of approximately 500 metres (Figure 20), aligning with domain-specific knowledge regarding the cooling effect of water bodies (Wu and Zhang, 2019). To further reduce overfitting risks associated with smaller local sample sizes, the hyperparameters for the local random forests were fixed as detailed in Table 4. Additionally, a histogram depicting the optimal number of neighbours was generated to assess variations in maximum distances across the local models.

Table 4: Hyperparameters for local random forests.

Parameter	Number of estimators	Maximum depth	Minimum samples per split	Minimum samples per leaf	Maximum features
Value	300	5	6	6	sqrt

To combine the global and local model predictions, the theoretically informed approach of Sun et al. (2024) was applied, since it considers the spatial autocorrelation of the dataset. The package PyGRF, which was developed for their study was employed to determine α .

3.4.3. GWRF model evaluation

To evaluate the model, several performance metrics were used: R^2 , MAE, MSE, and Mean Absolute Percentage Error (MAPE) (See Appendix B for more information). Additionally, scatter plots were made of the predicted and actual values, to investigate potential biases.

3.4.4. Interpretation of the relationship with LST

To interpret the model predictions, SHAP values were calculated. These were computed separately for the global and local models and then combined using the same weighting parameter (α) as for combining the global and local predictions. To interpret the resulting combined SHAP values, scatter plots were generated for each independent variable alongside their corresponding feature values. Additionally, the SHAP values were spatially visualized to explore geographic variation in feature importance and to identify which urban characteristics most strongly influenced LST across the study area.

4. Results

The results are presented according to the structure of the research questions and methodology. First, the results related to model fine-tuning and inference are described in detail. Next, a comparison of the different calculation methods for the VFs is given. Finally, the results regarding the relationship between urban characteristics and the LST are discussed.

4.1. Semantic segmentation

For semantic segmentation, the process of hyperparameter tuning is first represented, followed by an overview of the most important evaluation metrics and interpretation of interclass confusion from the classification, using confusion matrices.

4.1.1. Hyperparameter tuning

Hyperparameter tuning was applied to obtain the optimal parameters for model training. The initial learning rate was smaller for Model 1 compared to Model 2, and the decay rate higher for Model 1 compared to Model 2. This suggested that Model 2 benefited from a smaller learning rate and a smaller exponential decay (Table 5).

Table 5: Optimal hyperparameters for model excluding (Model 1) and including shrub-merged (Model 2).

Parameter	Optimal value Model 1	Optimal value Model 2
Initial Learning Rate	$1.1377 * 10^{-5}$	$1.0934 * 10^{-4}$
Decay Rate	0.8	0.649
Batch Size	2	2
Optimizer	Adam	Adam

4.1.2. Training and validation loss during fine-tuning

Loss was used as evaluation metric to be minimized during the fine-tuning process. The initial loss - denoted by Epoch -1 - was higher for Model 2 due to the inclusion of the *shrub-merged* class, which the model had not been trained on previously. For Model 1, training was stopped after 3 epochs due to an increasing validation loss for higher epochs, despite a decreasing training loss. In contrast, Model 2's training and validation loss plateaued after epoch 6, with

the training loss being slightly smaller, indicating only minimal further improvements. Therefore, fine-tuning for Model 2 was stopped after epoch 10 (Figure 12).

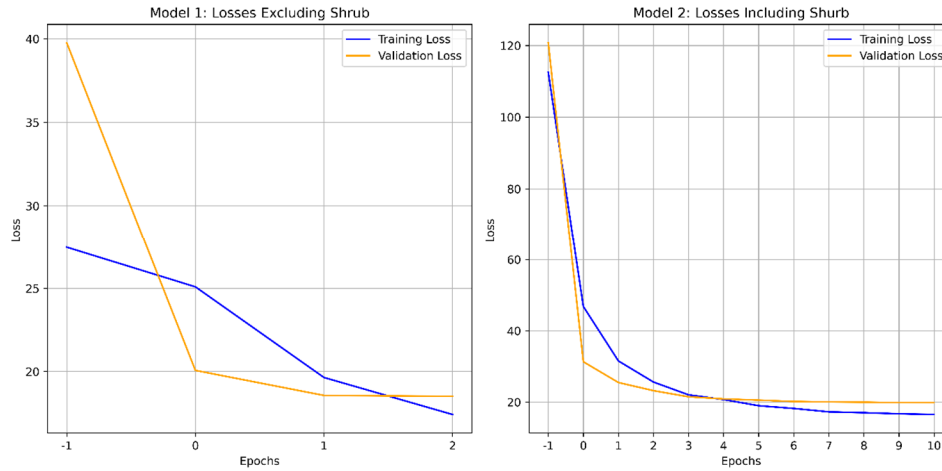


Figure 12: Model Losses during fine tuning for both training and validation dataset.

4.1.3. Model evaluation

For Model 1, all classes except *grass-merged* achieved scores above 90 for all evaluation metrics, indicating strong performance. The high precision and recall for these classes suggested a low number of both false positives and false negatives. *Grass-merged* also performed well, though its IoU, precision, and recall remained below 90. Precision was slightly higher than recall, suggesting that while the model was confident in identifying grass, it tended to miss some positive instances in the test data (Table 6). Most misclassifications for *grass-merged* were the result of ground truth labels being either predicted as *tree-merged* or *background-merged* (Figure 13).

For Model 2, the evaluation metrics for the classes *building-merged*, *sky-merged*, and *background-merged* were nearly identical to those of Model 1. However, some variation was observed in the vegetation classes. Specifically, all three evaluation metrics increased for the *grass-merged* class, indicating improved performance. In contrast, the metrics slightly decreased for the *tree-merged* class. For *shrub-merged*, precision was slightly higher than recall, suggesting higher confidence in predictions but a tendency to miss positive instances (Table 6). According to the confusion matrix, most misclassification of *grass-merged* occurred with *shrub-merged* and *background-merged*, although the overall confusion was lower than in Model 1 (Figure 13). For *shrub-merged*, misclassifications were primarily due to confusion with *tree-merged*. Conversely, the small number of misclassified *tree-merged* instances were mostly confused with *shrub-merged*.

Table 6: Evaluation metrics for models excluding and including shrub-merged.

Metric	tree-merged	grass-merged	shrub-merged	building-merged	sky-merged	background-merged
Model 1: Excluding shrub-merged						
IoU	0.93	0.79		0.92	0.98	0.96
Precision	0.96	0.89		0.96	0.99	0.98
Recall	0.97	0.87		0.96	0.99	0.98
Model 2: Including shrub-merged						

IoU	0.91	0.83	0.76	0.92	0.99	0.98
Precision	0.94	0.90	0.88	0.96	0.99	0.99
Recall	0.96	0.91	0.85	0.96	0.99	0.99

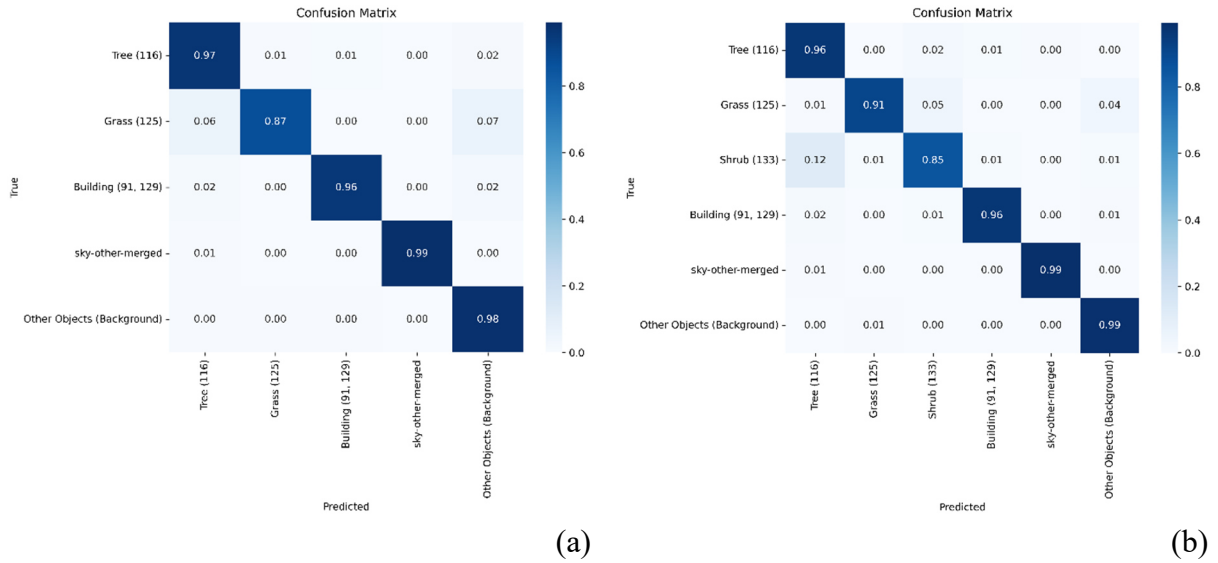


Figure 13: Confusion Matrices for (a) model 1 excluding shrub; and (b) model 2 including shrub.

4.2. View factor calculations

For calculating the VFs and continuation of the analysis, Model 2 - including *shrub-merged* - was used. The performance of *tree-merged* only marginally decreased, *grass-merged* improved, and the inclusion of *shrub-merged* enriched the model with additional information on the urban environment by including a more detailed description of vegetation.

4.2.1. Comparison of view factor calculation methods

Within this section, the three methods for calculating the VFs are presented, both in terms of the patterns they exhibited relative to each other and their impact on model behaviour.

4.2.1.1. Comparison between MVC and PIC

MVC consistently produced higher VF values for the categories *building-merged*, *grass-merged*, and *shrub-merged*. For these categories, the best-fit line lay below the identity line, and the slope was less than 1, indicating a positive bias relative to PIC. In contrast, *sky-merged* values were consistently higher for PIC, with the best-fit line positioned above the identity line. Additionally, *sky-merged* displayed a non-linear relationship with MVC. The best-fit line for *tree-merged* closely aligned with the identity line, suggesting strong agreement; however, the results were more dispersed around the line, as indicated by an R^2 value of 0.87 (Figure 14).

4.2.1.2. Comparison between MVC and Fisheye

The Fisheye method produced higher VF values for *tree-merged*, *building-merged*, and *sky-merged* compared to MVC. The best-fit lines for these categories deviated positively from the identity line, indicating a positive bias toward the Fisheye method. Despite this bias, the R^2 values were high, suggesting that most of the variance in the Fisheye imagery could be explained by MVC. As with the comparison between MVC and PIC, *sky-merged* again

exhibited a non-linear pattern. In contrast, *shrub-merged* and *grass-merged* showed a positive bias toward MVC, but with low R^2 values, indicating that only a small portion of the variance in the Fisheye data could be explained by MVC (Figure 15).

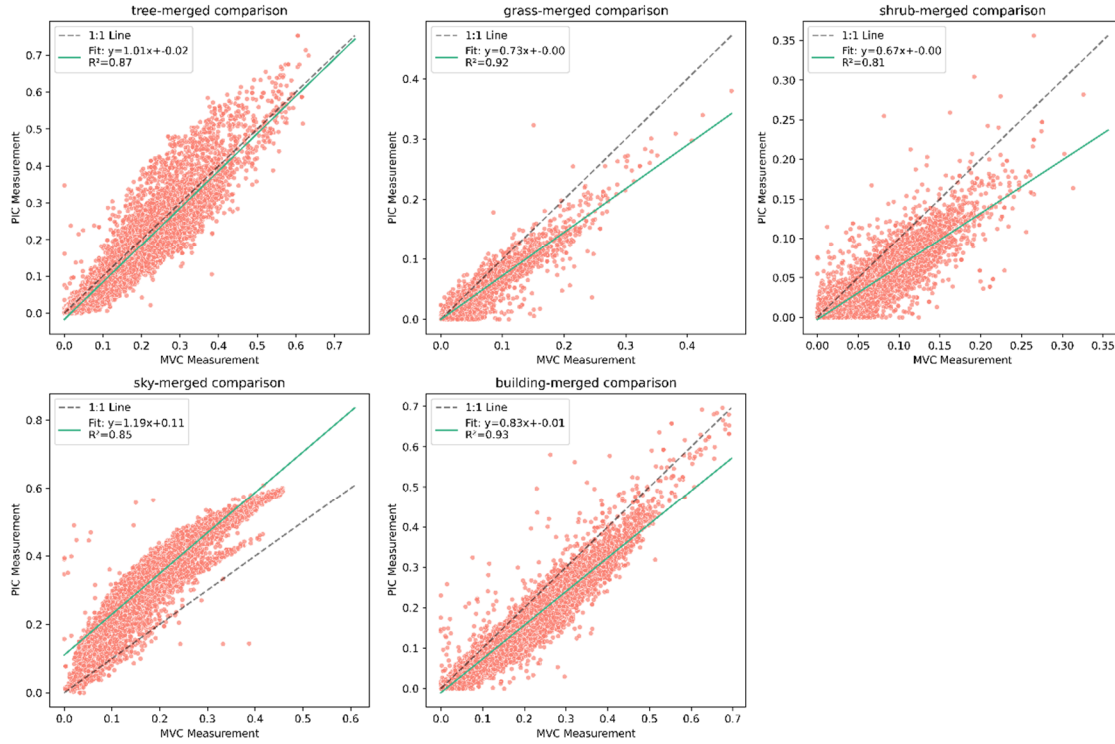


Figure 14: Comparison of VFs between MVC and PIC for the generalised categories.

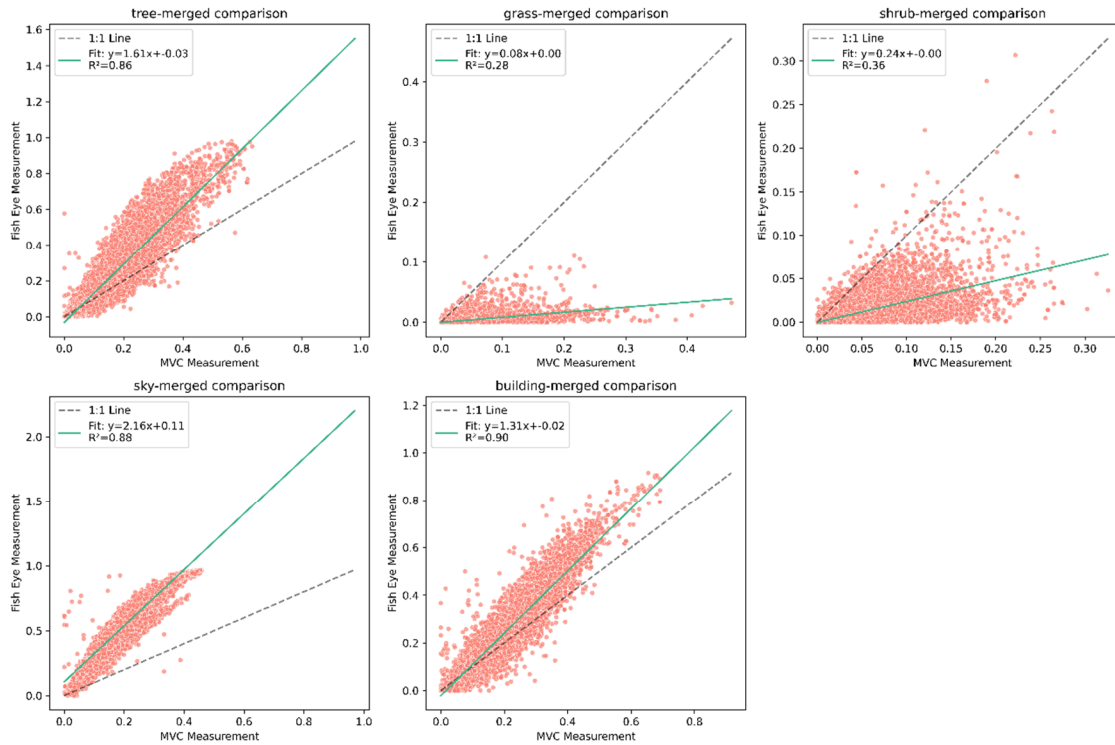


Figure 15: Comparison of VFs between MVC and Fisheye for the generalised categories.

4.2.1.3. Comparison between PIC and Fisheye

The strongest correspondence between PIC and Fisheye methods was observed for *tree-merged*, *sky-merged*, and *building-merged*. Once again, a positive bias toward the Fisheye method was apparent. However, the R^2 values for these categories were considerably higher, indicating that a larger portion of the variance in Fisheye measurements could be explained by PIC. For *grass-merged* and *shrub-merged*, a strong positive bias toward PIC was observed, and only a small proportion of the variance in Fisheye values was explained by PIC (Figure 16).

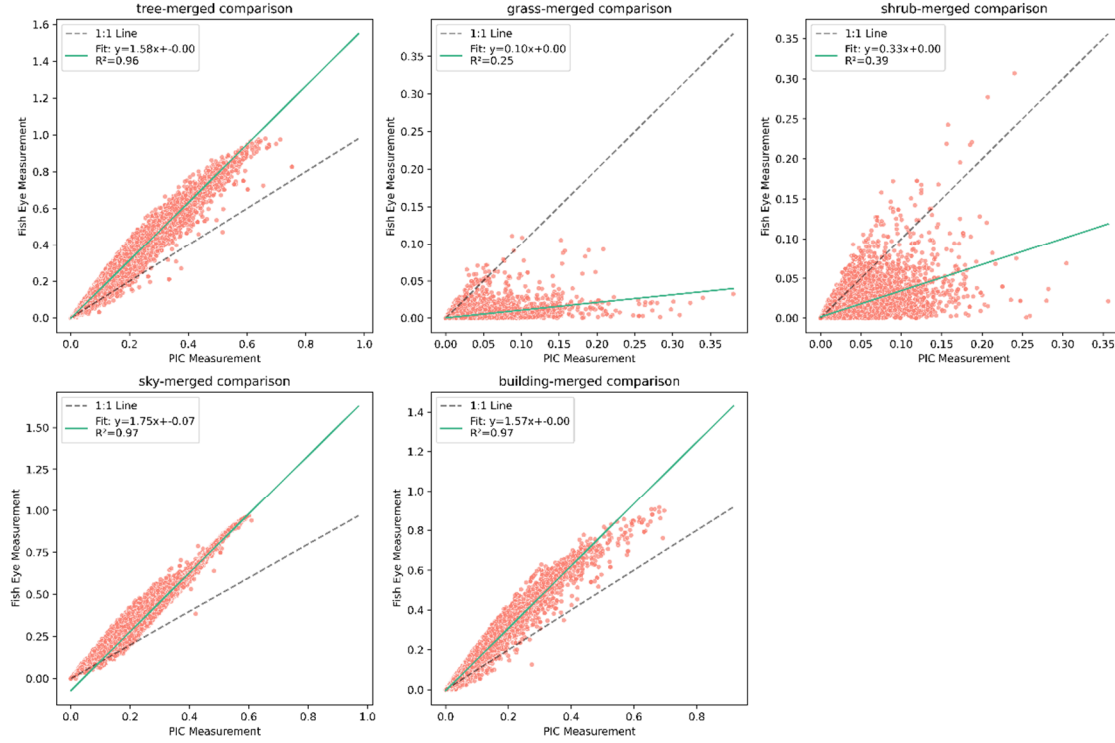


Figure 16: Comparison of VFs between PIC and Fisheye for the generalised categories.

4.2.2. Comparison of feature importances and model performance

The feature importances for the three models were very similar, with the overall pattern of influence remaining consistent across the different random forests. The most notable differences were observed for building-merged, where the highest feature importance was found in the PIC model, while the MVC and Fisheye models showed approximately the same (lower) values. For *tree-merged*, the highest value was observed in the MVC model, with PIC and Fisheye having similar, slightly lower values. However, all variations in feature importances across the methods were small relative to the magnitude of the feature importances between the different features (Figure 17).

In terms of evaluation metrics, the MVC model performed best, having the lowest MSE and MAE values and the highest R^2 . Nonetheless, as with feature importances, the differences between the models were minimal (Table 7). For the continuation of the analysis, the MVC method was used without any further comparisons with PIC and Fisheye. This was primarily done because the segmentation model was trained using the MVC images, and no objective evaluation was done for the other two methods.

Table 7: Evaluation metrics for random forest models for MVC, PIC and Fisheye.

Model	MSE ($^{\circ}\text{C}^2$)	MAE ($^{\circ}\text{C}$)	R ²
MVC	1.73	1.02	0.74
PIC	1.82	1.04	0.73
Fisheye	1.86	1.05	0.73

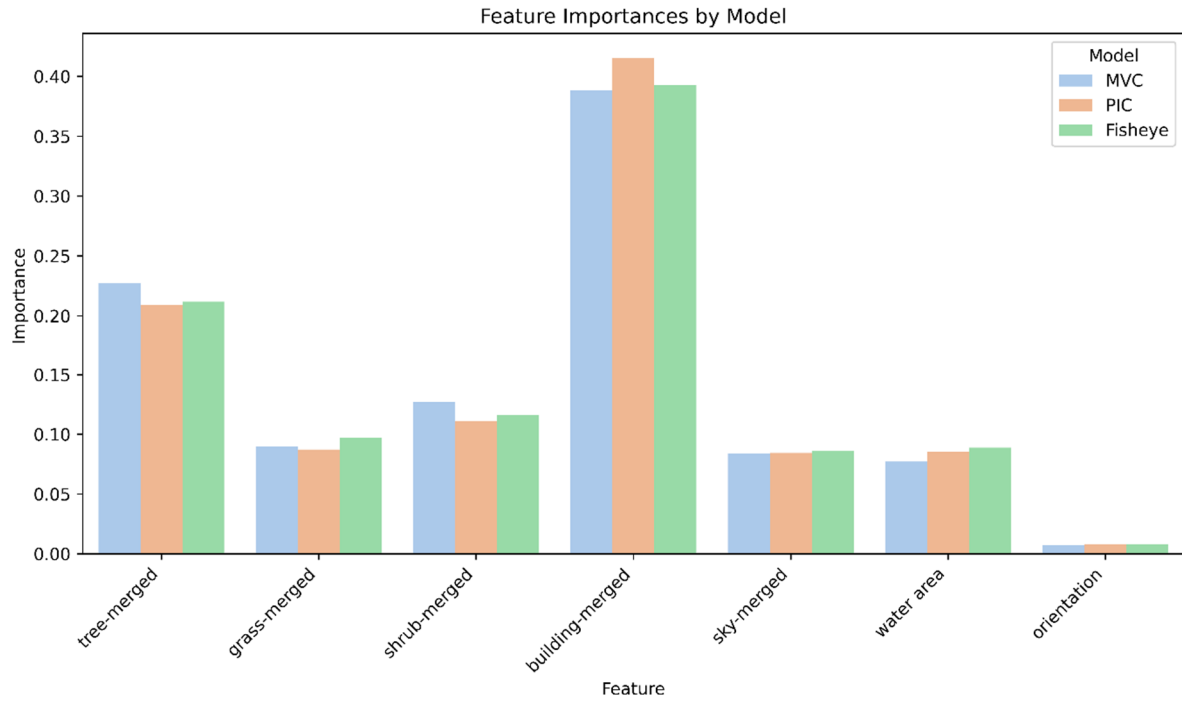


Figure 17: Feature importances across random forest models for MVC, PIC and Fisheye.

4.2.3. Spatial distribution of independent variables

Figure 18 illustrates the distribution of the different generalised VFs across Zurich. The VF *building-merged* clearly delineated the city centre, with higher values concentrated in the core and decreasing toward the outskirts. Conversely, *tree-merged* generally exhibited lower values in the centre and higher values further away. Both *grass-merged* and *shrub-merged* had generally low values across the municipality, though a slight increase was observed further from the city centre, particularly for *shrub-merged*. For *sky-merged*, the pattern was more subtle, with high values present throughout the entire city but showing a slight increase further from the city centre. For water area within 800m, the highest values were found close to Lake Zurich, and in the area directly north of the lake.

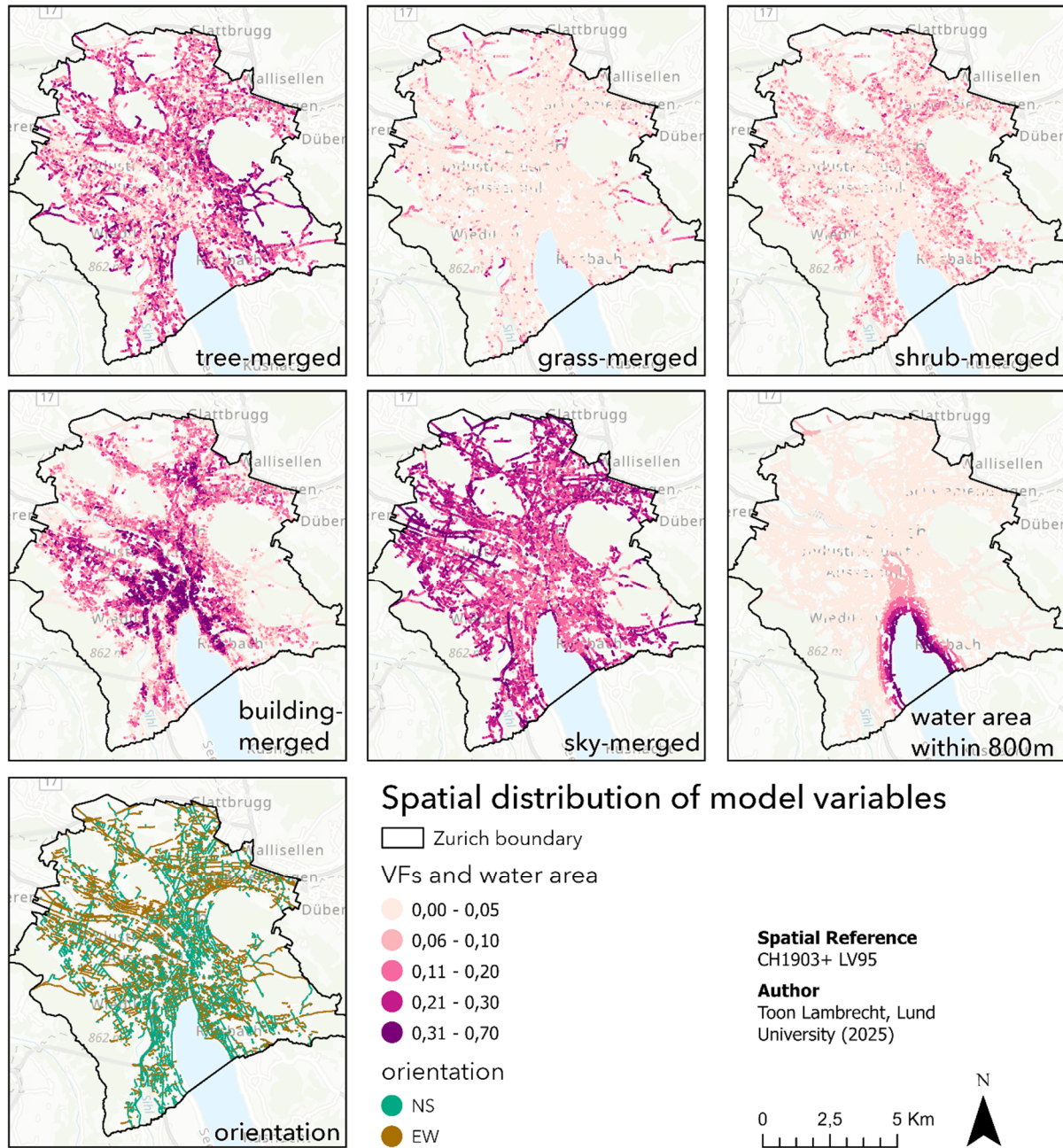


Figure 18: Spatial Distribution of view factors. All maps were individually included in Appendix C for easier interpretation.

4.3. Exploring the relationships between urban features and LST

GWRF was applied to explore the relationship between the urban features and LST. This included combining a global and local models. For the global model, hyperparameter tuning is presented. For the local models, bandwidth optimization is presented. Finally, the performance of the combined models is assessed along with the interpretation of the relationships found between the urban characteristics and LST, using SHAP analysis.

4.3.1. GWRF global model

During hyperparameter tuning, it became clear that the global model showed a tendency to overfit. All optimal hyperparameters were found at the edge of the defined intervals, each leaning toward configurations that increased the risk of overfitting (Table 8). Specifically, the depth had a largest defined value within the hyperparameter space, while the minimum samples per leaf and node, had the smallest value within the defined hyperparameter space.

Table 8: Result of hyperparameter tuning for the global model.

Parameter	Number of estimators	Maximum depth	Minimum samples per split	Minimum samples per leaf	Maximum features
Value	400	14	7	7	sqrt

4.3.2. GWRF local model

The restricted OOB optimization resulted in an optimal bandwidth of 200. Increasing the bandwidth beyond this value resulted in higher OOB errors, indicating no additional benefit from including more neighbours in the local models (Figure 19). The most frequently occurring maximum distance across all neighbour sets was 500 meters, although larger distances were also commonly observed. The minimum of the maximum distances was centred around 400 meters, suggesting that even for the densest regions, still a considerable area was considered for local model training (Figure 20).

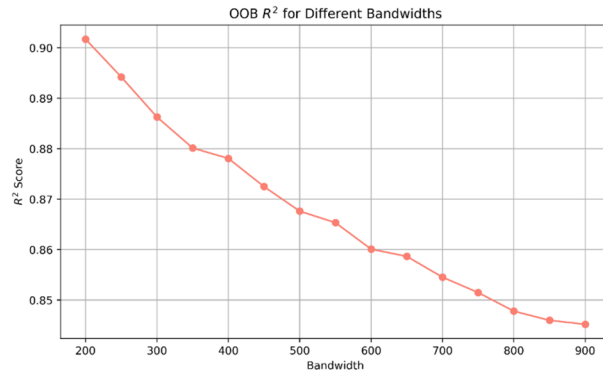


Figure 19: OOB bandwidth optimisation for GWRF.

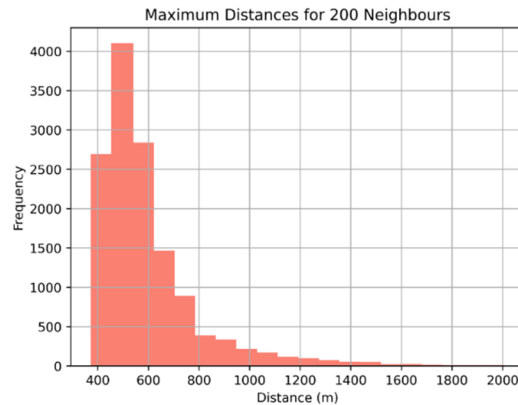


Figure 20: Maximum distance for every set of neighbours per instance. The distances to all neighbours were calculated for every instance, and the maximum distances were extracted.

4.3.3. GWRF combined model results

For combining the global and local models (Equation 1), an alpha value of 0.66 was found using the method of Sun et al. (2024). This resulted in good evaluation metrics for both training and testing datasets, suggesting good generalisation for the GWRF method. Most of the variance within both datasets was explained using the 7 variables, with R^2 values of 0.89 and 0.87 respectively. MAPE values for both datasets further suggested that both models performed well, since the average magnitude of the errors produced by the model was low (Table 9).

Table 9: Evaluation Metrics for Training and Tests datasets for combined models.

Dataset	R^2	MSE ($^{\circ}\text{C}^2$)	MAE ($^{\circ}\text{C}$)	MAPE (%)
Training	0.89	0.77	0.66	2.10
Testing	0.87	0.87	0.71	2.26

The prediction patterns for both the training and test datasets were similar. For lower LST values, the model showed a slight positive bias, with predicted values tending to lie above the identity line. In contrast, for higher LST values, the model showed a slight underestimation, as predictions were more frequently located below the identity line (Figure 21).

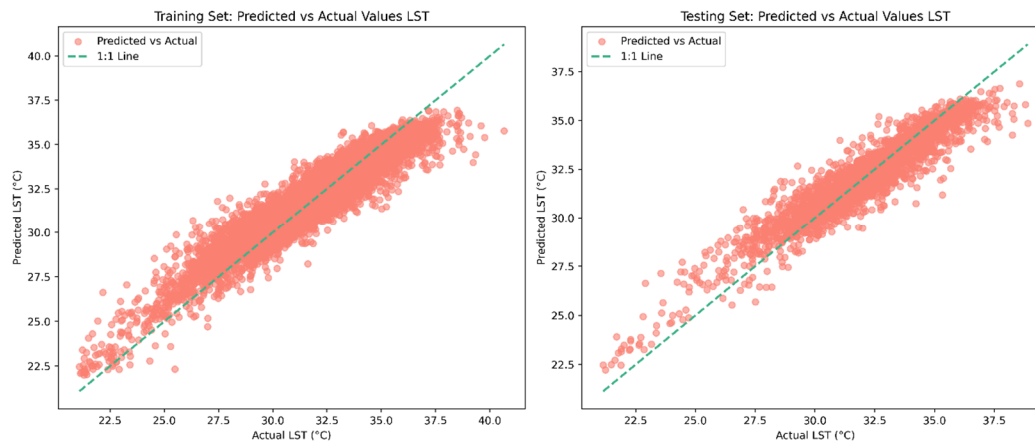


Figure 21: Scatter plot of predicted and actual LSTs for training and test datasets.

4.3.4. GWRF model interpretation

The SHAP values for *tree-merged* were almost consistently negative for VF values greater than 0.20, while predominantly positive SHAP values were observed for VFs below 0.10. Notably, SHAP values as low as -2°C were found for high VF values, and no strong non-linear patterns were apparent (Figure 22). The geographical distribution of the SHAP values for *tree-merged* exhibited two patterns. One was the overall urban pattern, with higher densities of more negative values towards areas outside of the city centre. Additionally, the city centre did contain localized negative values too, which coincided with large or dense street canopies. Furthermore, the green strip along the lake was clearly visible, with strong negative values. Positive SHAP values were mainly found inside denser areas coinciding with the absence of greenery or small canopies (Figure 23).

For *grass-merged*, SHAP values were almost consistently negative when VFs were below 0.10. However, there was no clear threshold at which SHAP values became consistently positive, as considerable variation was observed, including both positive and negative values even at low VFs. The magnitude of SHAP values for grass-merged was less pronounced than for tree-merged, ranging approximately from +0.60 to -1 (Figure 22). Looking at the geographical distribution, clearly visible were the grass patches in the south at the east side of the lake, and some larger patches in the outer areas of the study area. SHAP values within the centre were mostly slightly positive, falling within the interval of 0,06 - 0,50 (orange), or approximately zero (yellow). Only a few locations had moderate negative values (light green) (Figure 23).

In the case of *shrub-merged*, SHAP values were nearly always negative for VFs exceeding 0.05, while values gradually turned positive at lower VFs. A non-linear pattern was shown, with SHAP values flattening out early. As a result, higher VFs did not produce more negative SHAP values than mid-range ones. The range of SHAP values was mostly between +1 and -0.75 (Figure 22). The dense urban versus sparse urban pattern, like *tree-merged*, was also observed. However, the negative SHAP values were less pronounced but more abundant, especially in the outer areas of the study area. Some moderate negative values were also observed within the centre (Figure 23).

For *building-merged*, SHAP values were consistently positive for VFs greater than 0.20, while negative values were generally associated with VFs below 0.05. A clear non-linear pattern was observed, with SHAP values plateauing for VFs above 0.30, indicating no further increase in influence. The SHAP values for building-merged ranged mostly between -2 and +0.75 (Figure 22). The area directly north of the central lake as well as the area east of the lake had very high SHAP values. Moderately positive values were found throughout the whole study area, delineating the buildup areas, however, suburbs showed a tendency to lower positive or slight negative values. Furthermore, strips of negative values alongside fields or forest roads, coinciding with very low VFs, exhibited localized positive SHAP values for specific locations containing a higher VF (Figure 23).

In the case of *sky-merged*, SHAP values indicated a limited influence on model predictions across almost all VF levels. Most values were near zero, with a slight negative trend for very low VFs and a slight positive trend for very high VFs. However, the opposite was also observed for both VFs ranges, with small values showing higher SHAP values and vice versa. Overall, the impact remained moderate (Figure 22). No strong urban versus non-urban pattern was observed. Only the denser city area directly north of the lake showed positive SHAP values, while these locations coincided with lower VFs. A few smaller patches with strong positive values were also observed, these coincided with higher VFs. However, the opposite situation, high VFs with negative SHAP values was also observed (Figure 23).

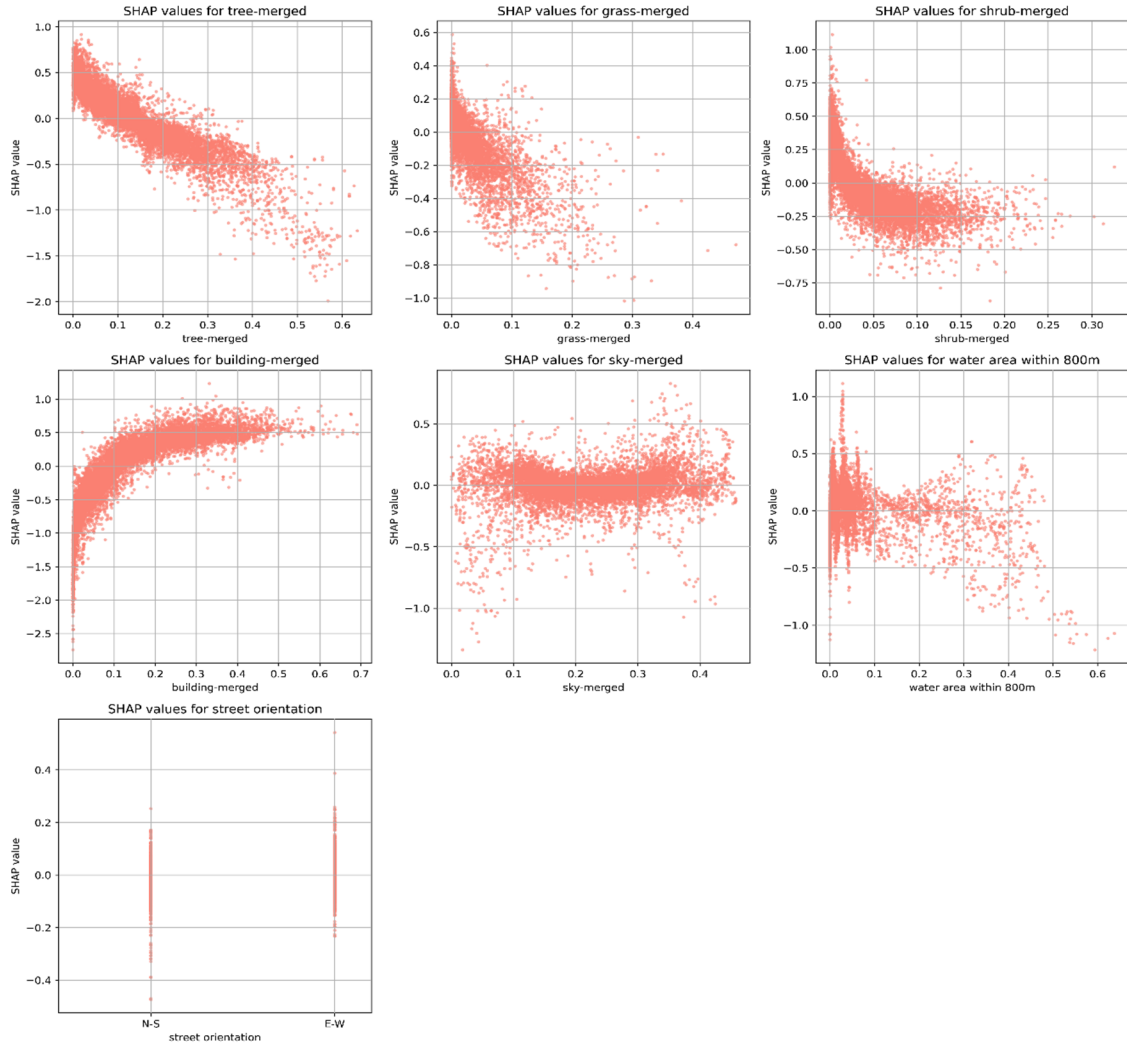


Figure 22: SHAP values for GWRf the 7 variables. Note, this figure concerned the SHAP values for the training dataset. The SHAP values for the test dataset were included in Appendix D. The scatter plot there exhibited the same overall pattern but was less dense due to the lower number of samples.

For *water area within 800 m*, SHAP values exhibited a noisy pattern for smaller areas, with both negative and positive values and a slight tendency towards positive values. For larger areas, the pattern became clearer, with mostly negative SHAP values, though many positive values still appeared. Only when the water area exceeded 0.5 did SHAP values become consistently negative. Overall, SHAP values for this variable ranged primarily between +1 and -1 (Figure 22). Strong negative values were found directly alongside the lake, coinciding with larger ratios. However, values with higher ratio's but not directly alongside the lake were also found. Furthermore, the noisy pattern observed in Figure 22, was also spatially visible with values approximating 0 in the north-north-east and northwest and moderately positive values for denser areas closer to the river that flows through the city. Moderate negative values were also found further away from the denser areas, while no stream were nearby (Figure 23).

Finally, for *street orientation*, SHAP values were similar for both E-W and N-S orientations, with a slight tendency towards positive values for E-W oriented streets and a slight tendency towards negative values for N-S orientations. The overall influence of street orientation on the model was small, with SHAP values mostly falling between -0.20 and +0.20 (Figure 22).

Almost all values were approximating zero (yellow), therefore, exhibiting no clear spatial trend. However, some localized streets deviating from the yellow class were found. Here, it seemed that streets N-S oriented overall had more negative values while streets E-W oriented had more positive values (Figure 23).

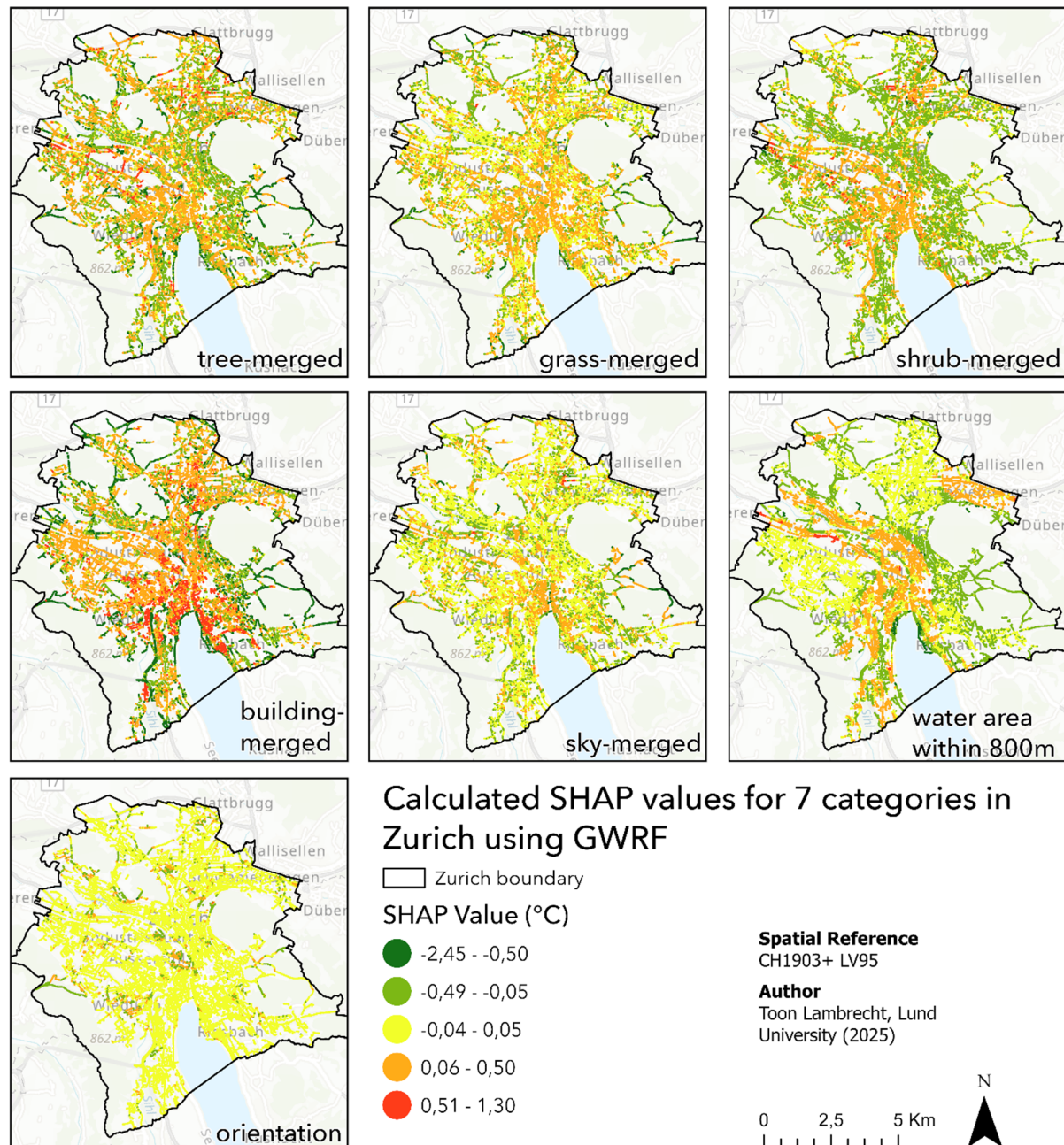


Figure 23: Geographical distribution of SHAP values for 7 categories, calculated using GWRF. Note, within these maps, all SHAP values - both training and testing - are included. All maps are individually included in Appendix E for easier interpretation.

5. Discussion

Within this study, three research questions were addressed. The first research question examined whether multiple aspects of the urban environment could be accurately mapped using GSV data through deep learning techniques. The second research question examined the differences between the VF calculation methods in terms of their interrelationships and influence on model behaviour. The third research question examined the impact of the different extracted urban features on LST. In the following sections, reflections are first offered on the data collection process. Each research question is then explicitly addressed, along with the associated methodological challenges, and additional context is provided for the interpretation of the results.

5.1. Data collection

This section reflects on the data collection process, focusing on the limitations encountered in using GSV and LST data. Challenges included spatial mismatches due to GSV positional uncertainty and potential biases introduced during LST downscaling.

5.1.1. Google Street View data collection

Although the method of Onáčillová et al. (2022) also allowed for retrieval of LST data at a 10 m resolution, this option was not used due to the positional uncertainty of the GSV data (50 m). Since this uncertainty exceeded the spatial resolution of the 10 m thermal data, the 30 m resolution was chosen instead. While this reduced the mismatch between GSV sample points and their corresponding LST values, it did not completely resolve this. For future research it would be recommended to change the API radius parameter from the default value of 50 m to a smaller value, to further reduce the possible locational mismatches. However, this will likely also result in lower data availability (*Google Maps Platform Documentation*, n.d.).

5.1.2. Temperature data collection

The upscaling approach of Onáčillová et al. (2022) for LST data introduced two limitations. The first concerns potential biases in the behaviour of the GWRF model and the associated SHAP values. The three indices used in the downscaling process - NDVI, NDBI, and NDWI - were also indirectly represented in the GSV-derived features (three vegetation classes, *building-merged*, and water area). Although these indices differed in both perspective (bird's-eye vs pedestrian) and calculation method (normalized indices for downscaling vs. VFs for GSV data), they conceptually extracted the same features which may have caused the random forest model to partially fit to the regression parameters used in the multiple linear regression. Despite this being difficult to prove, it should be kept in mind when interpreting the results and consider the possible biases in feature importances.

The second limitation was the lack of effective validation methods for evaluating the downscaled LST data. Ideally, higher-resolution LST data captured simultaneously with the sharpened LST data would allow for a robust assessment of the upscaling model's performance. However, such data are rarely available, and if they were, it would largely eliminate the need for downscaling altogether. In situ LST measurements, while an alternative, are not ideal for validation due to their limited spatial coverage. Collected at a much finer scale, they lack the spatial representativeness required to evaluate coarser-resolution remote sensing products

(Onačillová et al., 2022; Yu et al., 2011). Essentially, to independently and comprehensively study LST variations and their relationship with urban features at a city scale, higher-resolution, open-source thermal satellite data is necessary.

5.2. Semantic segmentation

Previous studies that used SV data to investigate the relationship between LST and greenery made limited distinctions between different types of greenery or included no other urban characteristics beyond greenery in their analyses (Li et al., 2022; Lu et al., 2023; Wei et al., 2024). To address research question 1, this study investigated whether the urban environment could be accurately described using a broader range of features for extraction, with the possibility of introducing additional classes by fine-tuning an existing model. It was found that, through fine-tuning with customized annotations, the model's performance improved, and an additional class could be successfully introduced. Further context regarding the fine-tuning process is provided in the following sections, along with a discussion of potential limitations.

5.2.1. Model selection

The architecture and model were chosen based on their reported performance, accessibility, and the availability of existing frameworks for model usage and fine-tuning. However, a more structured approach to model selection could have further improved the results of the semantic segmentation. First, the features of interest should be clearly defined in advance. Existing benchmark datasets could then be evaluated based on their relevance to these features. Second, once a suitable benchmark dataset is identified, different architectures trained on this benchmark dataset could be compared by annotating a small subset of the data, followed by an evaluation of the prediction results on this annotated subset.

5.2.2. Model training

During hyperparameter tuning, it was found that Model 2 benefited from a smaller learning rate and decay rate compared to Model 1. This may be attributed to the original classification scheme being retained in the Model 1, as it had already been trained on all categories and therefore required smaller parameter adjustments during the process. Additionally, Model 1 was trained for only 3 epochs, whereas the Model 2 was trained for 10 epochs. This further suggested that incorporating the new category required a longer training period to allow the model sufficient time to adjust its parameters effectively.

5.2.3. Model evaluation

Overall, the inclusion of *shrub-merged* in Model 2 enriched the models' representation of vegetation and shifted the nature of confusion between classes in a better direction. Although clear definitions were provided for the three vegetation types, confusion between *shrub-merged* and *tree-merged* was expected, especially in cases involving larger shrubs. However, this form of confusion was less problematic than that between *tree-merged* and *grass-merged*, since *shrub-merged* and *tree-merged* were more similarly defined, with both types including 'woody vegetation' into their definition. Furthermore, by introducing *shrub-merged*, the new model appeared to improve the semantic clarity and robustness of the classification overall based on the increased evaluation metrics for the other categories relative to Model 1.

5.2.4. Limitations

While the model benefited from customization to the dataset by creating image annotations, the process was labour intensive, since no pre-existing datasets were available. Additionally, the process of fine-tuning the model was time-consuming and computationally expensive due to the need for strong GPUs. The inclusion of *shrub-merged* only aggravated these problems due to the need for finer and more annotations and a longer model training time to effectively adjust the model parameters. Furthermore, the scalability of the fine-tuned model may be limited due to overfitting to the urban environment of Zurich itself, most likely necessitating further customization when applied in a different urban region.

To prevent overfitting, data augmentation techniques were applied. However, the quantity of annotations varied considerably across categories. For future research, it would be beneficial to incorporate data imbalance techniques to mitigate overfitting to the most prevalent categories. This becomes particularly important when extending the number of generalized VFs used to characterize the urban environment or when developing a model that aims at extending its applicability beyond a specific geographical area. Data imbalance techniques that could be implemented in future research include undersampling and oversampling, and cost-sensitive learning (Araf et al., 2024; Khushi et al., 2021)

5.3. View factor calculations

Previous studies employing VFs to characterise the urban environment in the context of urban heat modelling, primarily used three different methods: MVC, PIC and Fisheye. However, no comparison was yet made between these methods. Therefore, research question 2 was addressed by applying the three methods and comparing them both in terms of interrelationship and impact on model behaviour within the context of LST modelling. It was found that using the different methods resulted in different magnitudes and distributions in terms of their interrelationship. However, despite these differences, the influence on the model behaviour and performance was small, indicated by very similar feature importances and performance scores. Further context to this conclusion and the obtained results is provided within the following sections.

5.3.1. Comparison of VFs across MVC, PIC and Fisheye methods

The positive bias observed for all VFs - except for *sky-merged* - in MVC relative to PIC can likely be attributed to a certain degree of image overlap in the MVC method, due to its wider FOV compared to the heading angles. However, the overall correspondence between the methods remained relatively high for all categories. Greater discrepancies were observed between MVC and PIC on one hand, and the Fisheye method on the other for *grass-merged* and *shrub-merged*. Specifically, the Fisheye method resulted in noticeably lower values for these categories. Grasses and shrubs were both often located closer to the ground, resulting in their projection towards the periphery of the fisheye image, and consequently, lower VF values. The correspondence for *tree-merged*, *building-merged*, and *sky-merged* was higher, but a positive bias in the Fisheye method was observed relative to MVC and PIC, likely due to the larger size of these objects, which caused them to occupy more central regions of the image.

The non-linearity observed in *sky-merged* values between the PIC and Fisheye methods on one hand, and the MVC method on the other, may be attributed to differences in data representation. The panoramic images used for the PIC and Fisheye calculations were subject to distortions

due to the use of the equirectangular projection method. As a result of this projection technique, areas closer to the top and bottom of the image experience greater distortion, contributing to the observed non-linear patterns (Otto, 2023).

5.3.2. Influence of MVC, PIC and Fisheye methods on model behaviour

Given the comparable model behaviour and performance, the PIC method presents a computational advantage. As it relies on a single image projection to capture the entire hemisphere, the number of images required for prediction was significantly lower -13,801 for PIC compared to 165,612 for MVC. While the Fisheye method also relied on one image, the reprojection process was computationally expensive, since both images and predictions had to be reprojected. A downside of the PIC and Fisheye method may be that panoramic images pose challenges for semantic segmentation due to inherent distortions, which complicate feature extraction (Hu et al., 2022).

5.4. Exploring the relationships between urban features and LST

Previous studies that used SV data to investigate the relationship between LST and street greenery made limited distinctions between different types of greenery or included no other urban characteristics beyond greenery in their analyses (Li et al. 2022; Lu et al., 2023; Wei et al., 2024). Furthermore, no studies were found that addressed spatial non-stationarity effects simultaneously with non-linear regression techniques. Therefore, research question 3 was addressed by exploring the usage of GWRF integrated with SHAP analysis to examine the relationship between urban street features and LST. It was found that this method could effectively describe spatially non-stationary and non-linear effects, and that the obtained results were comparable with domain specific knowledge.

Within the following sections, more context is provided to the used modelling techniques. Additionally, the obtained results are compared with existing literature, to show how this workflow can be used to study the influence of urban features on LST at a street level.

5.4.1. OOB bandwidth optimisation

To apply the GWRF model, OOB optimisation was used to determine the optimal bandwidth. Since its introduction, this approach has been widely adopted across various fields with generally good results (Georganos & Kalogirou, 2022; Nduwayezu et al., 2023; Zhou et al., 2025). However, the method presents a specific challenge when applied to continuous datasets.

In OOB optimisation, predictions are generated using locally trained random forests, while the overall performance metric is computed across all samples. For continuous data, this setup tends to favour smaller bandwidths, even when these may lead to less stable local predictions. This bias arises because smaller bandwidths result in models trained on narrower data distributions leading to predictions made within a more limited range. While this can produce seemingly high accuracies, it increases the risk of overfitting.

To date, no studies were found that have directly addressed this issue from an optimisation perspective. However, Sun et al. (2024) proposed a new approach to bandwidth selection, in which the optimal bandwidth was derived based on the spatial scale at which spatial autocorrelation was most significant. Another approach could include a bandwidth sensitivity

analysis along with OOB optimisation; to analyse how variable the model parameters are for different bandwidths. This would offer an estimate of both model performance and stability.

5.4.2. Model performance

An overestimation of low LST values and an underestimation of high LST values were observed for the GWR model. The biases were small, but they were structural and have implications on the model interpretation. While it's difficult to exactly attribute these biases to specific model parameters, a few options are likely.

High density areas were associated with higher LSTs overall. Within these areas, an underestimation of LSTs may be attributed to an underestimation of the positive effect of high building VFs on LST or an overestimation of mitigating factors such as vegetation or water. Generally, higher mitigating VFs (greenery) were mostly associated with areas outside of the city centre, where lower LST values were observed overall. When similar high VFs were observed within the city centre, this may have led to an overestimation of the mitigating effect. Conversely, high building VFs were mostly associated with higher LSTs within the city centre. When higher building VFs were found in overall lower density areas, this may have led to an overestimation of low LST values. Using these explanations, the biases likely reflect the influence of the global model - that did not account for spatial non-stationarity - propagating through the final results.

Both are suggestions of possible influences. However, to fully understand these biases and their effects on the interpretation of the final SHAP values, a more precise analysis should be performed. This could include mapping the differences between model predictions and actual values and investigate the locations more closely where the biases were the largest. The SHAP values can then be used to understand the model predictions at these locations.

5.4.3. Theoretical reflections on SHAP analysis

While SHAP analysis offers a straightforward and intuitive way of interpreting complex machine learning models, it's important to be cautious when interpreting the actual values. SHAP analysis does not reveal how the real-world works, it explains how the model works, and more specifically how each feature at a given location contributed to a given prediction. This distinction comes down to the fundamental difference between correlation and causation: SHAP values capture correlations identified by the model but do not imply causal relationships with the dependent variable. For causal inference, specialized tools have been developed that account for potential confounding effects in model predictions, enabling the identification of causal relationships and supporting more informed policy decisions (Lundberg, 2021; Ma & Tourani, 2020). Within this research, no such tools were implemented, making it difficult to establish any causal relationships. However, it is possible to rely on domain specific knowledge and reflect how well the performed analysis reflected known relationships and compared to other studies focusing on the same features.

5.4.4. Practical interpretation of results

Overall, SHAP values ranged from approximately -2.5°C to $+2^{\circ}\text{C}$, with *tree-merged* and *building-merged* features having the strongest influence. In context, temperatures varied between approximately 21°C and 41°C within the study area (Figure 21). While this range of

SHAP values may appear relatively modest, it should be interpreted in the context of the spatial resolution of the analysis. The LST data had a resolution of 30 m, meaning that both the temperature variations captured by the model and the corresponding feature influences represented effects at that spatial scale.

The positive SHAP values for *tree-merged* for VFs smaller than 0.1, indicated a positive influence on the model predictions off small VFs on LST. This may be explained by lower *tree-merged* VFs being primarily found in denser regions with overall higher LST values. These low VFs, with positive SHAP values, could be understood as the relative absence of trees leading to higher temperature predictions, thus serving as an additional proxy of dense areas. Higher VFs on the other hand were consistently associated with strong negative values, both in dense and sparsely built areas, emphasizing the potential cooling benefits trees can offer in urban environments (Deng et al. 20219; Wu et al., 2024). Wei et al. (2024) found an overall relationship of a temperature decrease of 0.77 °C for every 0.1 increase in GVI. While no linear regression was performed in this study, using the SHAP values, the effect of tree-merged was overall less pronounced, but still within the same order of magnitude. However, direct comparison is difficult due the GVI capturing all vegetation, while *tree-merged* only includes tree like vegetation. Li et al. (2022) did distinguish between different types of vegetation but reported only individual regression coefficients and no specific temperatures, making direct comparisons difficult. They found values ranging from 0 to -4.88, though positive coefficients were not considered.

For *grass-merged*, the greater variation in SHAP values at very low VFs may be attributed to several factors. Smaller grass patches were distributed throughout the study area, and their influence likely depended on the local spatial context, such as urban density, surrounding surfaces, and ambient temperatures. These contextual differences may have contributed to both positive and negative model predictions at low VFs. However, for moderate to high VFs, grass consistently showed negative SHAP values, indicating a cooling effect. At first sight, this appears to contrast with the findings of Armson et al. (2022), who found that small grass patches had a stronger surface temperature cooling effect than larger ones, with differences of up to 24 °C for small patches compared to 15 °C for larger ones, based on in-situ measurements. This difference may be explained by differences in spatial scale: in-situ measurements reflect highly localized conditions, while satellite-based observations represent larger areas (here 30 m resolution). Small grass patches captured via satellite are likely influenced by surrounding surfaces, resulting in mixed effects and, in some cases, positive SHAP values. Li et al. (2022) reported regression coefficients as low as -7.36 for grass, although most values were much smaller, often ranging between 0 and -0.24, suggesting a more modest to small cooling effect overall.

For *shrub-merged*, positive contributions to LST were found at low VFs, with almost consistently negative contributions for VFs above 0.05. Like *grass-merged* and *tree-merged*, these positive contributions do not necessarily indicate a warming effect but may instead reflect the local spatial context. The relatively smaller cooling influence of shrubs compared to trees aligns with the findings of Yu et al. (2018), who reported a stronger negative association between taller vegetation and LST, suggesting shrubs are less effective in reducing temperatures. Li et al. (2022) similarly found the cooling effect of shrubs to be limited based on hotspot analysis, although they did report regression coefficients for *shrub-merged* as low

as -3.08. They also noted that shrub-related cooling effects were more evenly distributed across the city. This is consistent with the pattern observed in this study, where SHAP values for *shrub-merged* were generally less pronounced than for *tree-merged*, but more widely dispersed across the study area. Given that the SHAP values were more widely distributed and negatively associated with LST, could mean that shrubs can be used as cooling measure for locations where sufficient place lacks for trees. However, it should be noted that this analysis specifically concerns LST and that the negative contributions say nothing about other temperature metrics such as MRT, physiological equivalent temperature (PET), or shadow provision which are better estimates of thermal comfort (Kim et al., 2022; Klemm et al., 2015; Li & Ratti, 2018).

Negative SHAP values for *building-merged* at low VFs may be attributed to these lower VFs mostly occurring in low-density areas. Overall, higher VFs were associated with a positive contribution to LST, with the largest impacts observed in the city centre, where SHAP values reached up to 1.3 °C. This positive influence of higher *building-merged* VFs aligns with the findings of Zhang et al. (2019), who reported a positive correlation between high building VFs and LST. It also supports theoretical expectations related to street canyon geometry, which lowers the system's effective albedo due to radiative trapping. This geometry induces heat absorption and lowers heat dissipation, primarily because of multiple path reflections and limited sky exposure (Santamouris, 2011). Other studies have shown that high-rise buildings can have a cooling influence in lower-density areas (Chen et al., 2019; Yuan et al., 2021). While this effect was not directly examined in this study - high VFs in this case do not necessarily indicate building height, but rather high street density - no pattern was observed that would suggest a similar relationship here.

For *sky-merged*, the overall influence on LST was limited, with SHAP values approximating zero for medium-low to medium-high VFs. This relatively small influence contrasts with findings from previous research involving the SVF, where a stronger positive association with daytime MRT/LST was reported (Kim et al., 2022; Zhang et al., 2019). However, it is important to consider the timing of the thermal imagery within this study, which was captured between 10:00 and 10:20. At this hour, incoming solar radiation - and consequently surface heating - may still have been limited, potentially reducing the effect of SVF on LST in more open areas. Nonetheless, certain patterns were observed. Higher SHAP values were observed for low SVFs within denser urban areas, which may be attributed to reduced heat dissipation due to the urban canyon effect (Santamouris, 2011). In contrast, some low SVF values were associated with negative SHAP values. Considering their spatial distribution, these locations often coincided with areas having high *tree-merged* VFs, which are associated with lower temperatures overall.

The effects of water area within 800 m were noisy for small and medium-high values. This contrasts with the findings of Wu and Zhang (2019), who reported that Suzhou Bay had a mitigating effect on LST up to 800 m. However, their study focused exclusively on the Bay and did not include smaller water bodies such as rivers. Additionally, the area surrounding the Bay was more densely built compared to Zurich. This distinction is particularly relevant, as the spatially noisy pattern observed in this study appears to stem from the variable capturing latent differences in urban structure between high- and low-density areas. To better estimate the actual influence, it may have been beneficial to use a smaller buffer distance - more reflective of Zurich's city scale - and potentially to consider only the large lake, excluding smaller water streams such as the Limmat.

Street orientation had a limited overall effect on LST, with a slight positive bias observed for E-W oriented streets. Kim et al. (2022) found that E-W oriented streets were associated with higher MRT, which aligns with the findings of Zaki et al. (2020), who reported a positive influence of E-W orientation on PET. Although direct comparisons are difficult due to the use of different thermal metrics, the directionality of the effect appears consistent across studies. Furthermore, the modest influence observed in this study may be partly attributed to the spatial resolution of the LST data (30 m), which may not accurately capture the nuances of street orientation.

5.4.5. Methodological considerations

This was the first study implementing GWRF to examine LST in combination with VFs derived from GSV data and SHAP analysis. Using the assumption of spatial non-stationarity increased the models' performance and revealed spatial patterns that might otherwise remain uncaptured. SHAP analysis provided a straightforward way of interpreting the model's behaviour and spatial patterns. Challenges included proper bandwidth selection and determining the optimal weighting parameter. Furthermore, implementing SHAP analysis and GWRF is computationally expensive compared to traditional regression methods such as multiple linear regression and GWR. Performing comparisons across different settings required multiple runs, demanding computational power beyond standard CPUs, even with parallelisation.

5.5. Future research

In the future, the classes for feature extraction with semantic segmentation could be further extended with other relevant GSV extractable classes such as impervious surfaces or building albedo to further enhance the understanding of urban characteristics on LST (W. Liu et al., 2017; Zhang et al., 2019). Additionally, other variables such as building height and density could be included (Yuan et al., 2021). While density was implicitly accounted for using GWRF, a more explicit representation could provide insights into the exact effects.

Regarding temperature data, only daytime observations were used. However, a comparative study using nighttime images could provide additional insights into changing feature importances (Depietri et al., 2011; Zhang et al., 2019). Additionally, future higher resolution thermal data could provide an unbiased estimate of LST and reveal finer grained highly localized patterns of influence.

For future research regarding GWRF for LST estimation in combination with SHAP values, the overall robustness of the method should be tested using different case studies. Additionally, it could be explored how causal SHAP analysis methods could be included to account for confounding factors and extend the analysis beyond correlation methods.

6. Conclusions

In this study different methodologies were explored for mapping urban characteristics using GSV data and exploring the relationship with LST in the city of Zurich. It was found that trees, shrubs, grass, sky and buildings could be accurately mapped using an open-source deep learning framework. This shows the potential of SV data as a tool for describing the urban environment, offering an improved representation of the 3D urban form compared to traditional satellite imagery. It was also shown that the classification scheme of the model could be extended with

additional classes, for use-case customization and classification optimization through model fine-tuning. This shows that in future research, specific elements that are currently not included in semantic frameworks can be added to explore feature specific influences on LST, or other dependent variables of interest.

Three common methods for calculating VFs to characterize the urban environment were found in the literature (MVC, PIC and Fisheye). These methods were compared in terms of magnitude relative to each other, and in terms of influence on model performance. Relative to each other, non-linear effects and biases were observed, which were attributed to data representation. Nonetheless, the overall influences on the model behaviour were similar, with no method resulting in significantly better performance for predicting LST or big differences in feature importances. This indicates that for future research in the context of predicting LST, it would make sense to either make a theoretical argument for a certain method depending on the use case or choose the most computationally efficient method, which was PIC.

To study the relationship between LST and 3 types of greenery and other urban characteristics, a method was proposed that combined GWRF with SHAP analysis. It was found that this method effectively accounted for spatial non-stationarity and non-monotonous relationships between the different urban characteristics and LST, providing a more reliable estimate of relationships within the dataset than correlation analysis or GWR. Challenges that remain are proper bandwidth and weighting parameter selection. SHAP analysis revealed that buildings had the strongest positive effect on LST at moderate to high VF levels, displaying a clear non-linear trend. In contrast, trees had the most significant cooling effect within similar VF ranges. Shrubs showed a moderate yet more spatially distributed cooling effect, indicating their potential in areas where space constraints limit the use of large canopy trees. While these patterns were clearly identified through SHAP analysis, further research is needed to explore how causal SHAP values could be integrated into this framework. This could support more informed urban policy decisions aimed at mitigating the UHI effect and improving thermal comfort. Finally, the proposed framework should be tested further with other case-studies.

Declaration about the use of generative AI

During the preparation of this thesis, I used generative AI to help me in the writing process of some chapters, focusing on finding grammar and vocabulary errors. However, it was never used directly for text generation.

References

- Aletba, S. R. O., Hassan, N. A., Jaya, R. P., Aminudin, E., Mahmud, M. Z. H., Mohamed, A., & Hussein, A. A. (2021). Thermal performance of cooling strategies for asphalt pavement: A state-of-the-art review. *Journal of Traffic and Transportation Engineering (English Edition)*, 8(3), 356–373. <https://doi.org/10.1016/j.jtte.2021.02.001>
- Allen, R. G., Smith, M., Perrier, A., & Pereira, L. S. (1994). An update for the definition of reference evapotranspiration. *ICID bulletin*, 43(2), 1-34.
- Araf, I., Idri, A., & Chair, I. (2024). Cost-sensitive learning for imbalanced medical data: a review. *Artificial Intelligence Review*, 57(4). <https://doi.org/10.1007/s10462-023-10652-8>
- Armson, D., Stringer, P., & Ennos, A. (2012). The effect of tree shade and grass on surface and globe temperatures in an urban area. *Urban Forestry & Urban Greening*, 11(3), 245–255. <https://doi.org/10.1016/j.ufug.2012.05.002>
- Bhandari, R., Maithani, S., & Karnatak, H. (2022). A deep neural Network-Based approach for studying the relationship between land surface temperature and surface air temperature. *Journal of the Indian Society of Remote Sensing*, 50(3), 563–568. <https://doi.org/10.1007/s12524-021-01483-7>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/a:1010933404324>
- Chen, Y., Wu, J., Yu, K., & Wang, D. (2019). Evaluating the impact of the building density and height on the block surface temperature. *Building and Environment*, 168, 106493. <https://doi.org/10.1016/j.buildenv.2019.106493>
- Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2021). Masked-attention mask transformer for universal image segmentation. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2112.01527>
- Chugh, A. (2024, January 18). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? *Medium*. <https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>
- Cplusx. (2019). *GitHub - cplusx/google-street-view-panorama-download: A python tool to download google street view panorama images with given location*. GitHub. <https://github.com/cplusx/google-street-view-panorama-download>
- CVAT.ai. (n.d.). *CVAT Labeling Service | Expert data annotation for AI Projects*. <https://www.cvat.ai/annotation-services>
- Datamaro. (2024, September 30). PyPI. <https://pypi.org/project/datamaro/>
- Deng, J., Pickles, B. J., Kavakopoulos, A., Blanas, T., Halios, C. H., Smith, S. T., & Shao, L. (2019). Concept and methodology of characterising infrared radiative performance of urban trees using tree crown spectroscopy. *Building and Environment*, 157, 380–390. <https://doi.org/10.1016/j.buildenv.2019.04.056>
- Depietri, Y., Renaud, F. G., & Kallis, G. (2011). Heat waves and floods in urban areas: a policy-oriented review of ecosystem services. *Sustainability Science*, 7(1), 95–107. <https://doi.org/10.1007/s11625-011-0142-4>
- Fahmy, A. H., Abdelfatah, M. A., & El-Fiky, G. (2023). Investigating land use land cover changes and their effects on land surface temperature and urban heat islands in Sharqiyah Governorate, Egypt. *The Egyptian Journal of Remote Sensing and Space Science*, 26(2), 293–306. <https://doi.org/10.1016/j.ejrs.2023.04.001>
- GeeksforGeeks. (2024, September 25). *R Squared | Coefficient of determination*. GeeksforGeeks. <https://www.geeksforgeeks.org/r-squared/>
- Georganos, S., Grippa, T., Gadiaga, A. N., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., & Kalogirou, S. (2019). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, 36(2), 121–136. <https://doi.org/10.1080/10106049.2019.1595177>
- Georganos, S., & Kalogirou, S. (2022). A Forest of Forests: a spatially weighted and computationally efficient formulation of geographical random forests. *ISPRS*

- International Journal of Geo-Information*, 11(9), 471.
<https://doi.org/10.3390/ijgi11090471>
- Georgiou, A. M., & Akçit, N. (2017). Comparison of MODIS-derived land surface temperature with air temperature measurements. *Proceedings of SPIE - the International Society for Optical Engineering*, 55. <https://doi.org/10.1117/12.2279339>
- Ghaddar, N. Z., Ghali, N. K., & Ghaddar, N. N. (2017). The impact of the Air-Conditioning systems on the urban microclimate of Beirut City. *Renewable Energy and Power Quality Journal*, 15(7). <https://doi.org/10.24084/repqj15.504>
- Gong, F., Zeng, Z., Zhang, F., Li, X., Ng, E., & Norford, L. K. (2018). Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Building and Environment*, 134, 155–167. <https://doi.org/10.1016/j.buildenv.2018.02.042>
- Google Maps Platform Documentation. (n.d.). Google for Developers. <https://developers.google.com/maps/documentation/streetview>
- Guo, H., Ferrara, M., Coleman, J., Loyola, M., & Meggers, F. (2020). Air temperature and mean radiant temperature data, collected and simulated across a radiantly-heated high-bay laboratory. *Data in Brief*, 30, 105192. <https://doi.org/10.1016/j.dib.2020.105192>
- Guo, J., Han, G., Xie, Y., Cai, Z., & Zhao, Y. (2020). Exploring the relationships between urban spatial form factors and land surface temperature in mountainous area: A case study in Chongqing city, China. *Sustainable Cities and Society*, 61, 102286. <https://doi.org/10.1016/j.scs.2020.102286>
- HannaMergui. (2023a, June 22). MaskFormer: Per-Pixel Classification is Not All You Need for Semantic Segmentation. *Medium*. <https://medium.com/@HannaMergui/maskformer-per-pixel-classification-is-not-all-you-need-for-semantic-segmentation-1e2fe3bf31cb>
- HannaMergui. (2023b, July 10). MaskFormer2 : Masked-Attention Mask Transformer for universal image segmentation. *Medium*. <https://medium.com/@HannaMergui/maskformer2-masked-attention-mask-transformer-for-universal-image-segmentation-c3d14c546d6b>
- Haque, I. R. I., & Neubert, J. (2020). Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked*, 18, 100297. <https://doi.org/10.1016/j.imu.2020.100297>
- Hu, X., An, Y., Shao, C., & Hu, H. (2022). Distortion Convolution module for semantic segmentation of panoramic images based on the Image-Forming principle. *IEEE Transactions on Instrumentation and Measurement*, 71, 1–12. <https://doi.org/10.1109/tim.2021.3139710>
- Hua Ge, P. (2013). An infrared sphere method to measure mean radiant temperature. *ASHRAE Transactions*, 119, 1F.
- Johnson, G. T., & Watson, I. D. (1984). The Determination of View-Factors in Urban Canyons. *Journal of Climate and Applied Meteorology*, 23(2), 329–335. <http://www.jstor.org/stable/26181335>
- Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, 9, 109960–109975. <https://doi.org/10.1109/access.2021.3102399>
- Kim, E., Yun, S., Park, C., Heo, H., & Lee, D. (2022). Estimation of mean radiant temperature in urban canyons using Google Street View: a case study on Seoul. *Remote Sensing*, 14(2), 260. <https://doi.org/10.3390/rs14020260>
- Klemm, W., Heusinkveld, B. G., Lenzholzer, S., & Van Hove, B. (2015). Street greenery and its physical and psychological impact on thermal comfort. *Landscape and Urban Planning*, 138, 87–98. <https://doi.org/10.1016/j.landurbplan.2015.02.009>
- Koomen, E., & Diogo, V. (2015). Assessing potential future urban heat island patterns following climate scenarios, socio-economic developments and spatial planning strategies. *Mitigation and Adaptation Strategies for Global Change*, 22(2), 287–306. <https://doi.org/10.1007/s11027-015-9646-z>
- Kundu, D., & Pandey, A. K. (2020). World Urbanisation: Trends and Patterns. In *Developing National Urban Policies* (pp. 13–49). https://doi.org/10.1007/978-981-15-3738-7_2

- Le, M. T., Le, H. T., Shukurov, I., & Slesarev, M. (2020). Sulfur-extended asphalt concrete with assessing the surface temperature of roads affecting urban heat island. *IOP Conference Series Materials Science and Engineering*, 869(2), 022007. <https://doi.org/10.1088/1757-899x/869/2/022007>
- Lee, J. S., Kim, J. T., & Lee, M. G. (2013). Mitigation of urban heat island effect and greenroofs. *Indoor and Built Environment*, 23(1), 62–69. <https://doi.org/10.1177/1420326x12474483>
- Li, B., Xing, H., Cao, D., Yang, G., & Zhang, H. (2022). Exploring the effects of roadside vegetation on the urban thermal environment using street view images. *International Journal of Environmental Research and Public Health*, 19(3), 1272. <https://doi.org/10.3390/ijerph19031272>
- Li, H. (2016). Impacts of pavement strategies on human thermal comfort. In *Elsevier eBooks* (pp. 281–306). <https://doi.org/10.1016/b978-0-12-803476-7.00013-1>
- Li, X., & Ratti, C. (2018). Mapping the spatial distribution of shade provision of street trees in Boston using Google Street View panoramas. *Urban Forestry & Urban Greening*, 31, 109–119. <https://doi.org/10.1016/j.ufug.2018.02.013>
- Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban Forestry & Urban Greening*, 14(3), 675–685. <https://doi.org/10.1016/j.ufug.2015.06.006>
- Liu, J., Zhang, S., Yin, L., Li, B., & Zhang, B. (2023). Exploring the relationship between seasonal variations of land surface temperature and urban morphological factors in complex urban areas. *Environmental Science and Pollution Research*, 30(21), 59861–59876. <https://doi.org/10.1007/s11356-023-26781-0>
- Lu, Y., Ferranti, E. J. S., Chapman, L., & Pfrang, C. (2023). Assessing urban greenery by harvesting street view data: A review. *Urban Forestry & Urban Greening*, 83, 127917. <https://doi.org/10.1016/j.ufug.2023.127917>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Lundberg, S. (2021, May 17). Be careful when interpreting predictive models in search of causal insights. *Medium*. <https://medium.com/data-science/be-careful-when-interpreting-predictive-models-in-search-of-causal-insights-e68626e664b6>
- Ma, S., & Tourani, R. (2020, August 19). *Predictive and Causal Implications of using Shapley Value for Model Interpretation*. PMLR. <https://proceedings.mlr.press/v127/ma20a>
- Mask2Former. (n.d.). Hugging Face. https://huggingface.co/docs/transformers/model_doc/mask2former
- Meili, N., Manoli, G., Burlando, P., Carmeliet, J., Chow, W. T., Coutts, A. M., Roth, M., Velasco, E., Vivoni, E. R., & Fatichi, S. (2020). Tree effects on urban microclimate: Diurnal, seasonal, and climatic temperature differences explained by separating radiation, evapotranspiration, and roughness effects. *Urban Forestry & Urban Greening*, 58, 126970. <https://doi.org/10.1016/j.ufug.2020.126970>
- Millison, A. (2019, August 1). *Appendix C: Koppen Geiger Classification Descriptions*. Pressbooks. [https://open.oregonstate.edu/permaculturedesign/back-matter/koppen-geiger-classification-descriptions/#:~:text=Cfb%20%3D%20Temperate%20oceanic%20climate%3B%20cold est,C%20\(50%20%C2%B0F\)](https://open.oregonstate.edu/permaculturedesign/back-matter/koppen-geiger-classification-descriptions/#:~:text=Cfb%20%3D%20Temperate%20oceanic%20climate%3B%20cold est,C%20(50%20%C2%B0F))
- Mohajerani, A., Bakaric, J., & Jeffrey-Bailey, T. (2017). The urban heat island effect, its causes, and mitigation, with reference to the thermal properties of asphalt concrete. *Journal of Environmental Management*, 197, 522–538. <https://doi.org/10.1016/j.jenvman.2017.03.095>
- Naeem, S., Cao, C., Qazi, W., Zamani, M., Wei, C., Acharya, B., & Rehman, A. (2018). Studying the Association between Green Space Characteristics and Land Surface Temperature for Sustainable Urban Environments: An Analysis of Beijing and Islamabad. *ISPRS International Journal of Geo-Information*, 7(2), 38. <https://doi.org/10.3390/ijgi7020038>
- Nduwayezu, G., Zhao, P., Kagoyire, C., Eklund, L., Bizimana, J. P., Pilesjo, P., & Mansourian, A. (2023). Understanding the spatial non-stationarity in the relationships between malaria

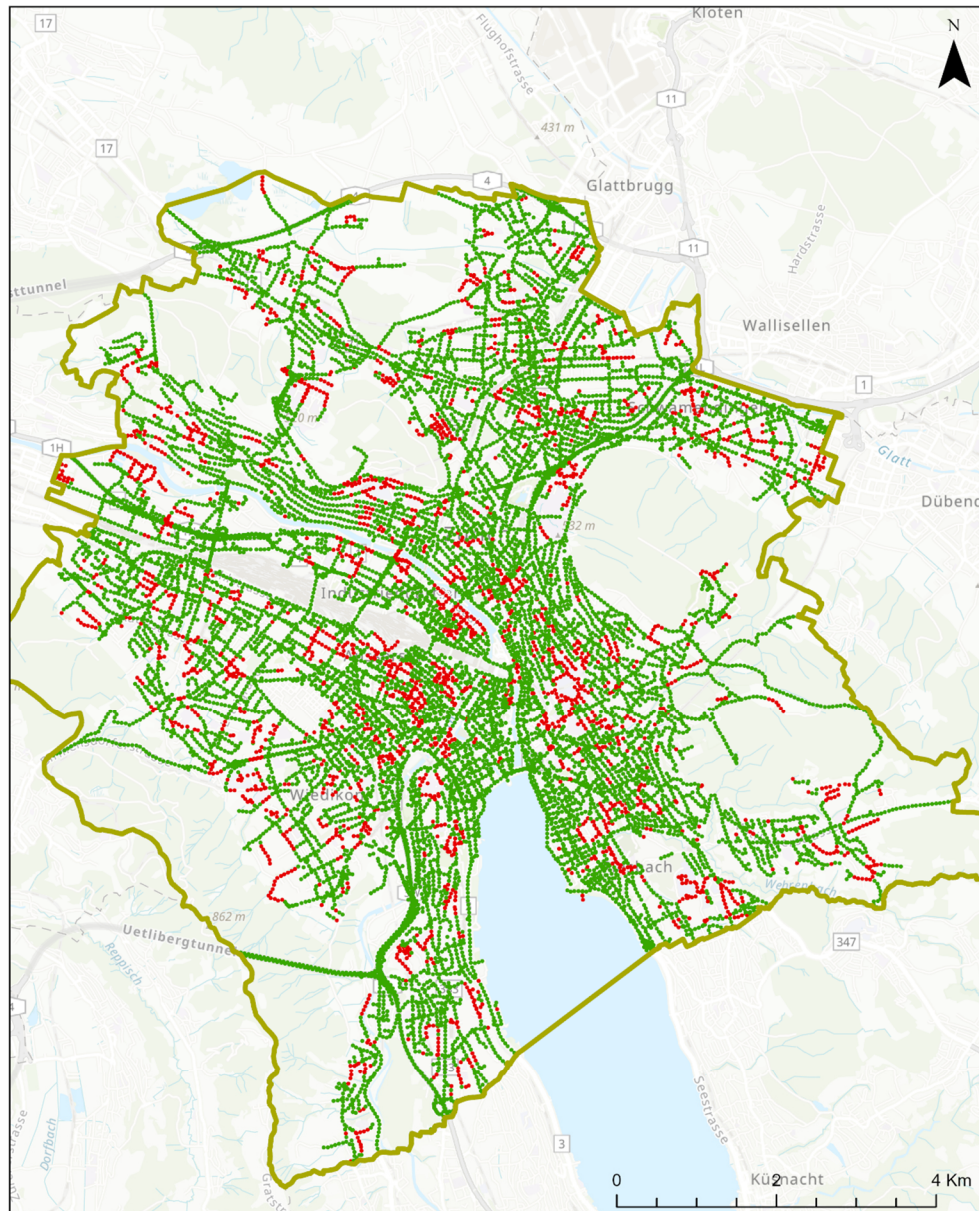
- incidence and environmental risk factors using Geographically Weighted Random Forest: A case study in Rwanda. *Geospatial Health*, 18(1). <https://doi.org/10.4081/gh.2023.1184>
- Oke, T. R. (1982). The energetic basis of the urban heat island. *Quarterly Journal of the Royal Meteorological Society*, 108(455), 1–24. <https://doi.org/10.1002/qj.49710845502>
- Onáčillová, K., Gallay, M., Paluba, D., Péliová, A., Tokarčík, O., & Laubertová, D. (2022). Combining Landsat 8 and Sentinel-2 data in Google Earth Engine to derive higher resolution land surface temperature maps in urban environment. *Remote Sensing*, 14(16), 4076. <https://doi.org/10.3390/rs14164076>
- Otto, Y. K. H. (2023, September 19). Equirectangular convolution in Panoramic image - Yu Kai Him Otto - Medium. *Medium*. <https://medium.com/@OttoYu/equirectangular-convolution-in-panoramic-image-147ed9bef96b>
- Owen, L. (2022). *Hyperparameter Tuning with Python: Boost Your Machine Learning Model's Performance Via Hyperparameter Tuning*. *Papers with Code - Semantic Segmentation*. (n.d.). <https://paperswithcode.com/task/semantic-segmentation>
- Precision-Recall*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html
- Roberts, A. (2023, February). *Mean Absolute Percentage Error (MAPE): What You Need To Know*. [https://arize.com/blog-course/mean-absolute-percentage-error-mape-what-you-need-to-know/#:~:text=point%20in%20time\).-Mean%20absolute%20percentage%20error%20measures%20the%20average%20magnitude%20of%20error,and%20the%20actuals%20is%2020%25](https://arize.com/blog-course/mean-absolute-percentage-error-mape-what-you-need-to-know/#:~:text=point%20in%20time).-Mean%20absolute%20percentage%20error%20measures%20the%20average%20magnitude%20of%20error,and%20the%20actuals%20is%2020%25)
- Sailor, D. J. (2010). A review of methods for estimating anthropogenic heat and moisture emissions in the urban environment. *International Journal of Climatology*, 31(2), 189–199. <https://doi.org/10.1002/joc.2106>
- Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest algorithm Overview. *Deleted Journal*, 2024, 69–79. <https://doi.org/10.58496/bjml/2024/007>
- Santamouris, M. (2011). *Energy and climate in the urban built environment*. Routledge.
- Sayad, B., Helmi, M. R., Osra, O. A., Abed, A. M., & Alhubashi, H. H. (2024). Microscale Investigation of Urban Heat Island (UHI) in Annaba City: Unveiling factors and mitigation strategies. *Sustainability*, 16(2), 747. <https://doi.org/10.3390/su16020747>
- Subedi, N., Zhang, L., & Zhen, Z. (2018). Bayesian geographically weighted regression and its application for local modeling of relationships between tree variables. *iForest - Biogeosciences and Forestry*, 11(5), 542–552. <https://doi.org/10.3832/for2574-011>
- Sun, K., Zhou, R. Z., Kim, J., & Hu, Y. (2024). PyGRF: an improved Python geographical random forest model and case studies in public health and natural disasters. *Transactions in GIS*. <https://doi.org/10.1111/tgis.13248>
- The Dataset object*. (n.d.). <https://huggingface.co/docs/datasets/v2.1.0/access>
- The Topographic Landscape Model TLM*. (n.d.). <https://www.swisstopo.admin.ch/en/information-topographic-landscape-model>
- United Nations, Department of Economic and Social Affairs, Population Division (2019). *World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420)*. New York: United Nations.
- Wang, X., Li, Z., Ding, S., Sun, X., Qin, H., Ji, J., & Zhang, R. (2023). Study on the Relationship between Urban Street-Greenery Rate and Land Surface Temperature Considering Local Climate Zone. *International Journal of Environmental Research and Public Health*, 20(4), 3294. <https://doi.org/10.3390/ijerph20043294>
- Wang, Y., Bakker, F., De Groot, R., Wortche, H., & Leemans, R. (2015). Effects of urban trees on local outdoor microclimate: synthesizing field measurements by numerical modelling. *Urban Ecosystems*, 18(4), 1305–1331. <https://doi.org/10.1007/s11252-015-0447-7>
- Wei, S., He, Z., Zhai, W., Zhao, C., & Li, Y. (2024). How does vegetation influence surface temperature across various road types and urban morphology types? *Building and Environment*, 112511. <https://doi.org/10.1016/j.buildenv.2024.112511>
- Wikipedia contributors. (2025, May 17). *Zurich*. Wikipedia. <https://en.wikipedia.org/wiki/Zurich>

- Wu, D., Zhang, Y., & Xiang, Q. (2023). Geographically weighted random forests for macro-level crash frequency prediction. *Accident Analysis & Prevention*, 194, 107370. <https://doi.org/10.1016/j.aap.2023.107370>
- Wu, Z., Shi, Y., Ren, L., & Hang, J. (2024). Scaled outdoor experiments to assess impacts of tree evapotranspiration and shading on microclimates and energy fluxes in 2D street canyons. *Sustainable Cities and Society*, 108, 105486. <https://doi.org/10.1016/j.scs.2024.105486>
- Wu, Z., & Zhang, Y. (2019). Water bodies' cooling effects on urban land daytime surface temperature: Ecosystem Service reducing Heat Island Effect. *Sustainability*, 11(3), 787. <https://doi.org/10.3390/su11030787>
- Xiao, Z., Chai, T., Li, N., Shen, X., Guan, T., Tian, J., & Li, X. (2024). Research advances in deep learning for image semantic segmentation techniques. *IEEE Access*, 12, 175715–175741. <https://doi.org/10.1109/access.2024.3496723>
- Yang, M., Ma, C., Zhao, S., Zhao, F., & Guo, J. (2024). Aiding and Opposing Cooling the Street Canyon through Trees and Water Bodies: Conflicting Effects from Temperature Reduction and Ambient Humidification. *Building and Environment*, 112513. <https://doi.org/10.1016/j.buildenv.2024.112513>
- Yu, Q., Acheampong, M., Pu, R., Landry, S. M., Ji, W., & Dahigamuwa, T. (2018). Assessing effects of urban vegetation height on land surface temperature in the City of Tampa, Florida, USA. *International Journal of Applied Earth Observation and Geoinformation*, 73, 712–720. <https://doi.org/10.1016/j.jag.2018.08.016>
- Yu, Y., Tarpley, D., Privette, J. L., Flynn, L. E., Xu, H., Chen, M., Vinnikov, K. Y., Sun, D., & Tian, Y. (2011). Validation of GOES-R Satellite land surface temperature Algorithm using SURFRAD ground measurements and statistical estimates of error properties. *IEEE Transactions on Geoscience and Remote Sensing*, 50(3), 704–713. <https://doi.org/10.1109/tgrs.2011.2162338>
- Yuan, B., Zhou, L., Dang, X., Sun, D., Hu, F., & Mu, H. (2021). Separate and combined effects of 3D building features and urban green space on land surface temperature. *Journal of Environmental Management*, 295, 113116. <https://doi.org/10.1016/j.jenvman.2021.113116>
- Zaki, S. A., Toh, H. J., Yakub, F., Saudi, A. S. M., Ardila-Rey, J. A., & Muhammad-Sukki, F. (2020). Effects of roadside trees and road orientation on thermal environment in a tropical city. *Sustainability*, 12(3), 1053. <https://doi.org/10.3390/su12031053>
- Zhang, Y., Middel, A., & Turner, B. L. (2019). Evaluating the effect of 3D urban form on neighborhood land surface temperature using Google Street View and geographically weighted regression. *Landscape Ecology*, 34(3), 681–697. <https://doi.org/10.1007/s10980-019-00794-y>
- Zhou, Y., Wei, G., Wang, Y., Wang, B., Quan, Y., Wu, Z., Liu, J., Bian, S., Li, M., Fan, W., & Dai, Y. (2025). Estimating regional forest carbon density using remote sensing and geographically weighted random forest models: A case study of Mid- to High-Latitude forests in China. *Forests*, 16(1), 96. <https://doi.org/10.3390/f16010096>

Appendices

Appendix A: Availability of Google Street View images in Zurich

This map shows the availability for GSV images for every sample point. Using a metadata request, it was checked for every location whether GSV imagery was available.



Zurich Google Street View Availability

- Zurich boundary
- Location Availability
 - Ok
 - Zero results

Spatial Reference
Name: CH1903+ LV95
GCS: GCS CH1903+
Author
Toon Lambrecht, Lund
University (2025)

Appendix B: Evaluation metrics

Intersection over Union (IoU) measures the overlap between the predicted and ground truth labels. It is calculated as the area of intersection divided by the total combined area (union) and is given by Equation 9:

$$IoU = \frac{\text{Area of Intersection}}{\text{Area of Union}} \quad (9)$$

Values closer to 1 indicate higher performance, while values closer to 0 indicate lower performance (Haque & Neubert, 2020).

Precision (P) represents the fraction of correctly classified positive pixels among all positively classified pixels. It is defined using Equation 10:

$$P = \frac{TP}{TP + FP} \quad (10)$$

Where TP is the number of true positives, and FP is the number of false positives. Consequently, a precision value close to 1 indicates that the model is highly confident in its predictions for a particular class, meaning that most predicted positives are indeed correct (Precision-Recall, n.d.).

Recall (R) represents the fraction of correctly classified positive pixels among all pixels that should have been positively classified. It is defined using Equation 11:

$$R = \frac{TP}{TP + FN} \quad (11)$$

Where FN is the number of false negatives. Consequently, a recall value close to 1 means that the model successfully detects most of the actual positive instances, while a lower value suggests that many true positives were missed (Precision-Recall, n.d.).

The **coefficient of determination (R^2)** quantifies the extent to which the variance in dependent variable is accounted for by the independent variable(s) and is calculated using Equation 12:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (12)$$

Where:

- n is the number of samples,
- y_i is the i th observation,
- $\hat{y}_i$ is the predicted value of y_i ,
- $\bar{y}$ is the mean value of all samples.

Essentially, it measures how much better the model predicts each points value compared to the mean. Values closer to 1 indicate a larger portion of the variance is explained, while values closer to 0 indicate only a low portion of the variance is explained (GeeksforGeeks, 2024).

The **Mean Absolute Error (MAE)** measures sum of absolute errors, divided by the sample size and is given by Equation 13:

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (13)$$

Where:

- n is the number of samples,
- y_i is the i th observation,
- $\hat{y}_i$ is the predicted value of y_i .

This error metric uses the same scale as the data of interest, and consequently, its interpretation is scale dependent (Chugh, 2024).

The **Mean Squared Error (MSE)** measures the average of the squares of the model errors and is given by Equation 14:

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \quad (14)$$

Where:

- n is the number of samples,
- y_i is the i th observation,
- $\hat{y}_i$ is the predicted value of y_i .

This error metric uses the squared scale of the data of interest, and consequently, its interpretation is scale dependent (Chugh, 2024).

The **Mean Absolute Percentage Error (MAPE)** represents the average magnitude of the error produced by the model and is given by Equation 15:

$$MAPE = \frac{\sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} * 100}{n} \quad (15)$$

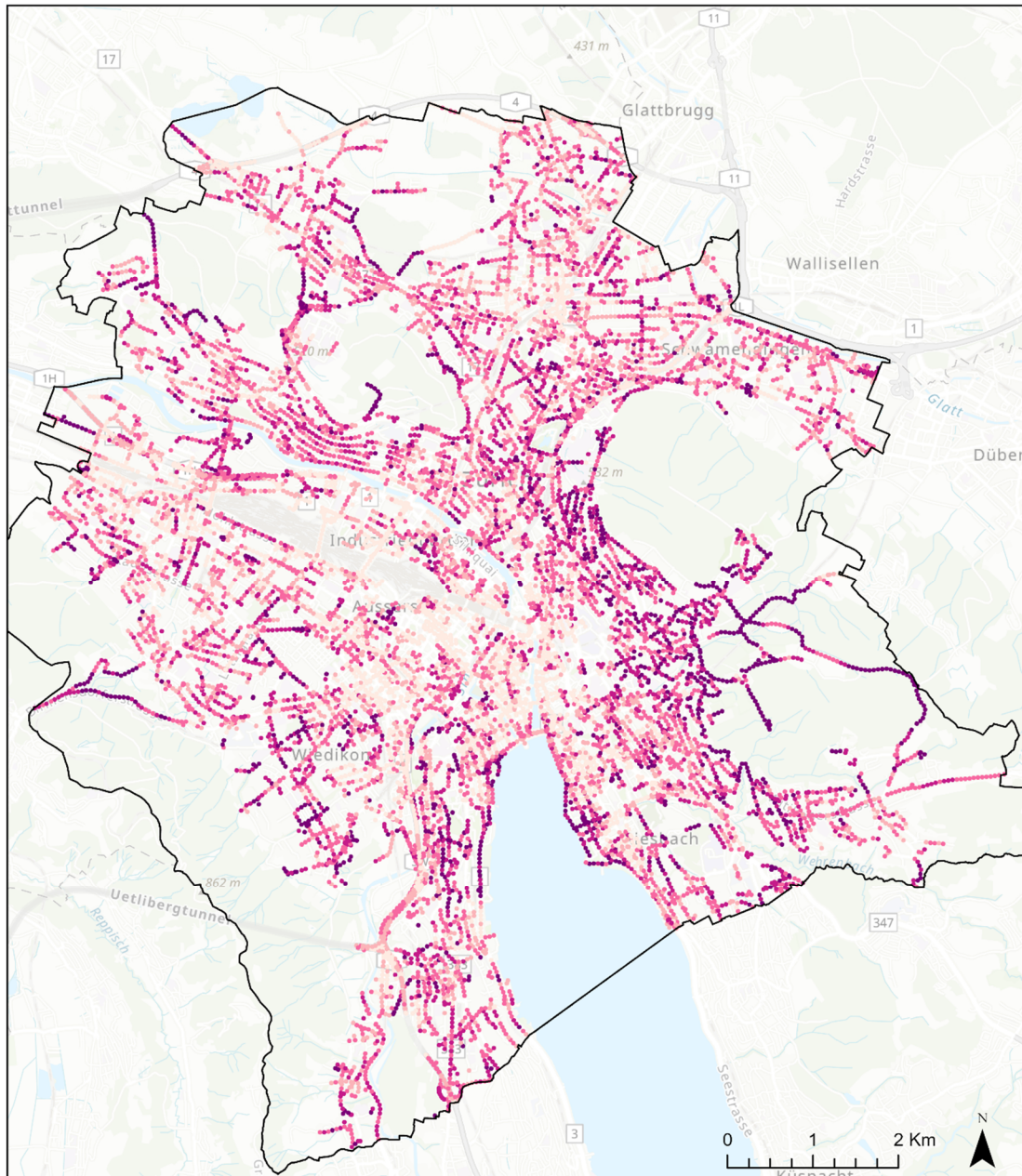
Where:

- n is the number of samples,
- y_i is the i th observation,
- $\hat{y}_i$ is the predicted value of y_i .

This metric provides an estimate of how far predictions are off on average. Generally, values below 10% are considered good. However, this also depends on the variability in the data and the range of the actual values (Roberts, 2023).

Appendix C: Individual maps of distribution of model variables

This appendix includes the individual maps that are shown together in Figure 18. They were individually included to aid interpretation. They show the spatial distribution of calculated VFs and the two additional drivers of temperature.



Spatial distribution of view factors for tree-merged - Zurich

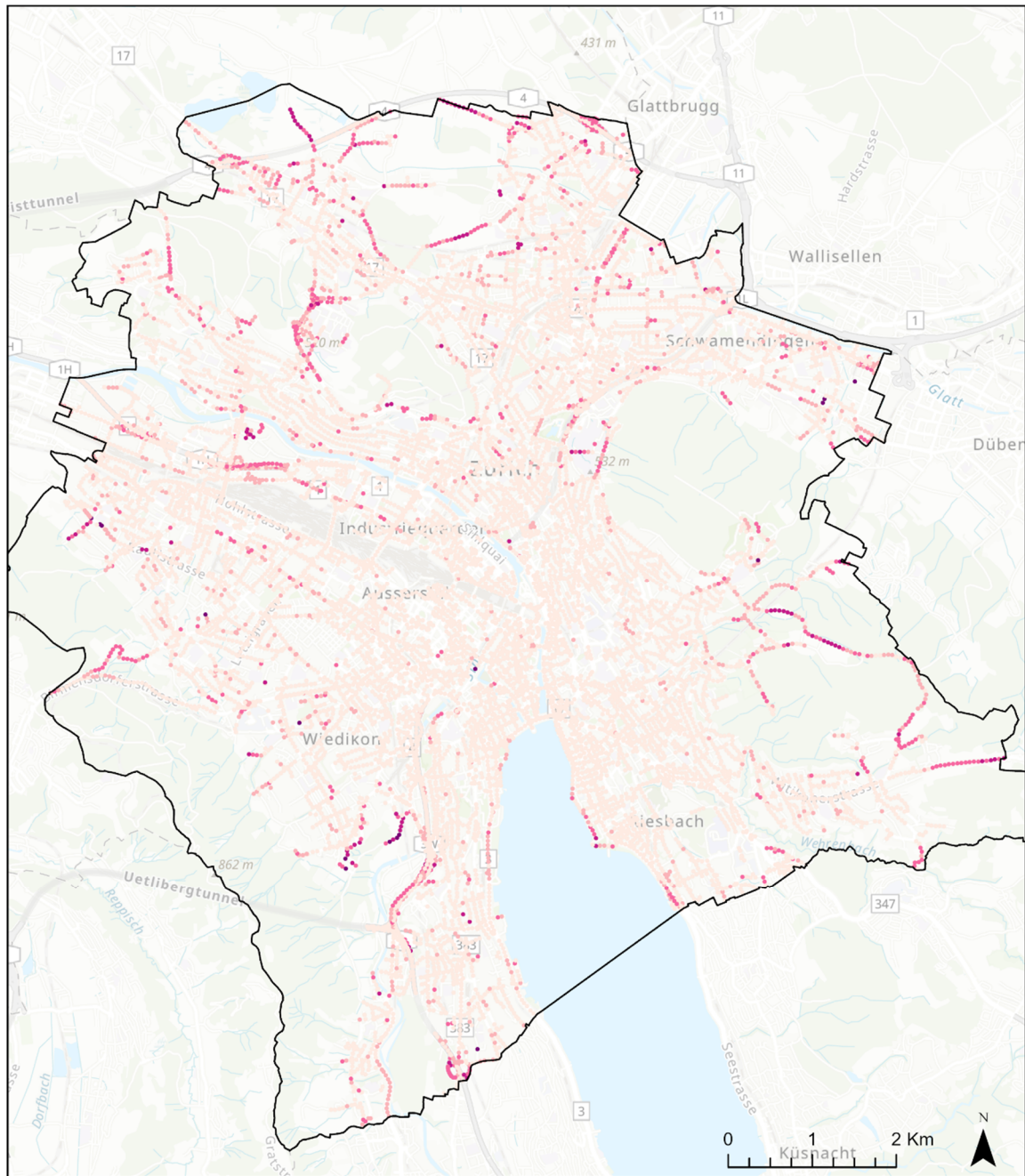
□ Zurich boundary

view factor

- 0,00 - 0,05
- 0,06 - 0,10
- 0,11 - 0,20
- 0,21 - 0,30
- 0,31 - 0,70

Spatial Reference
WGS 1984 UTM Zone 32N

Author
Toon Lambrecht, Lund
University (2025)



Spatial distribution of view factors for grass-merged - Zurich

□ Zurich boundary

view factor

0,00 - 0,05

0,06 - 0,10

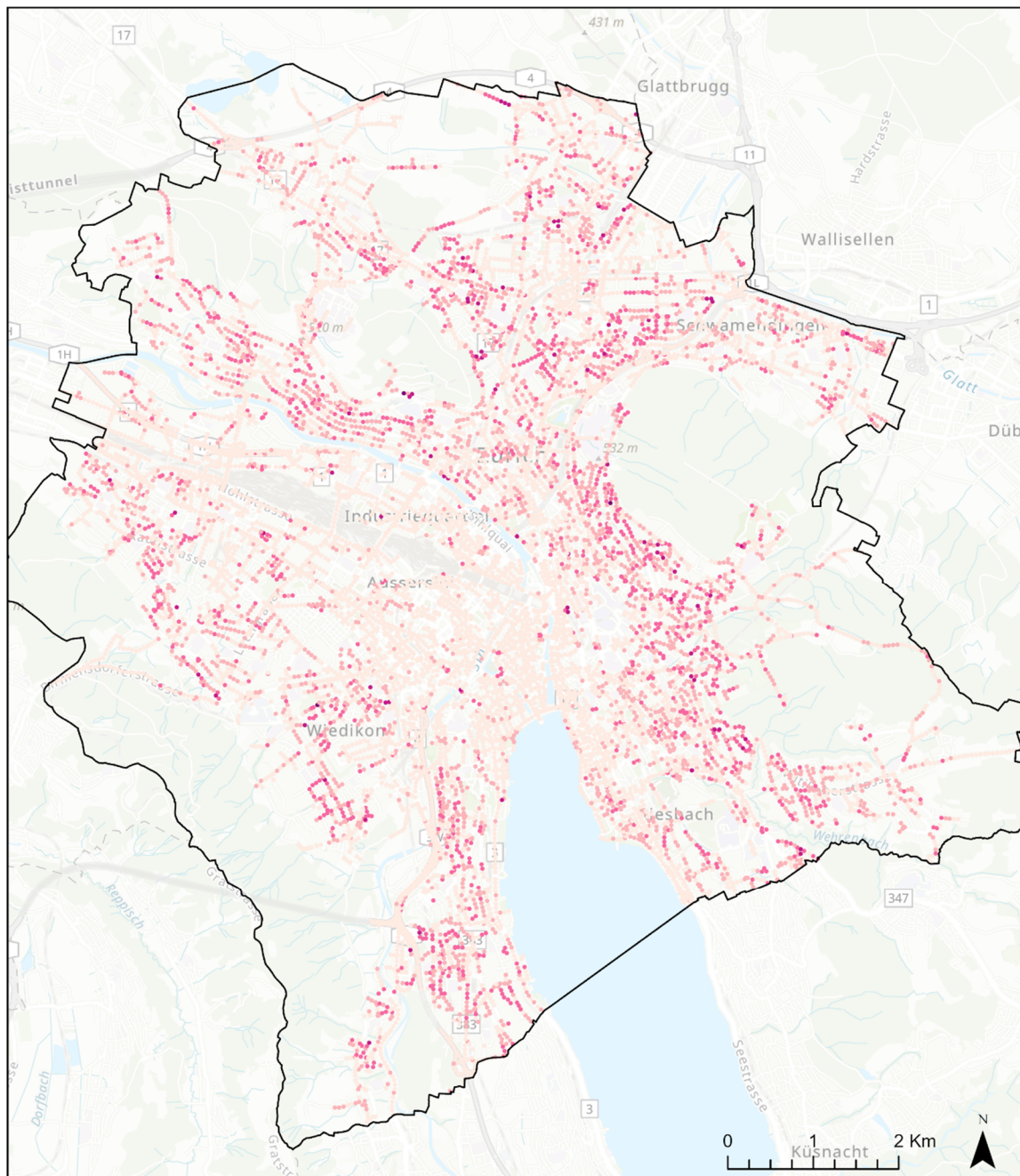
0,11 - 0,20

0,21 - 0,30

0,31 - 0,70

Spatial Reference
WGS 1984 UTM Zone 32N

Author
Toon Lambrecht, Lund
University (2025)



Spatial distribution of view factors for shrub-merged - Zurich

□ Zurich boundary

view factor

0,00 - 0,05

0,06 - 0,10

0,11 - 0,20

0,21 - 0,30

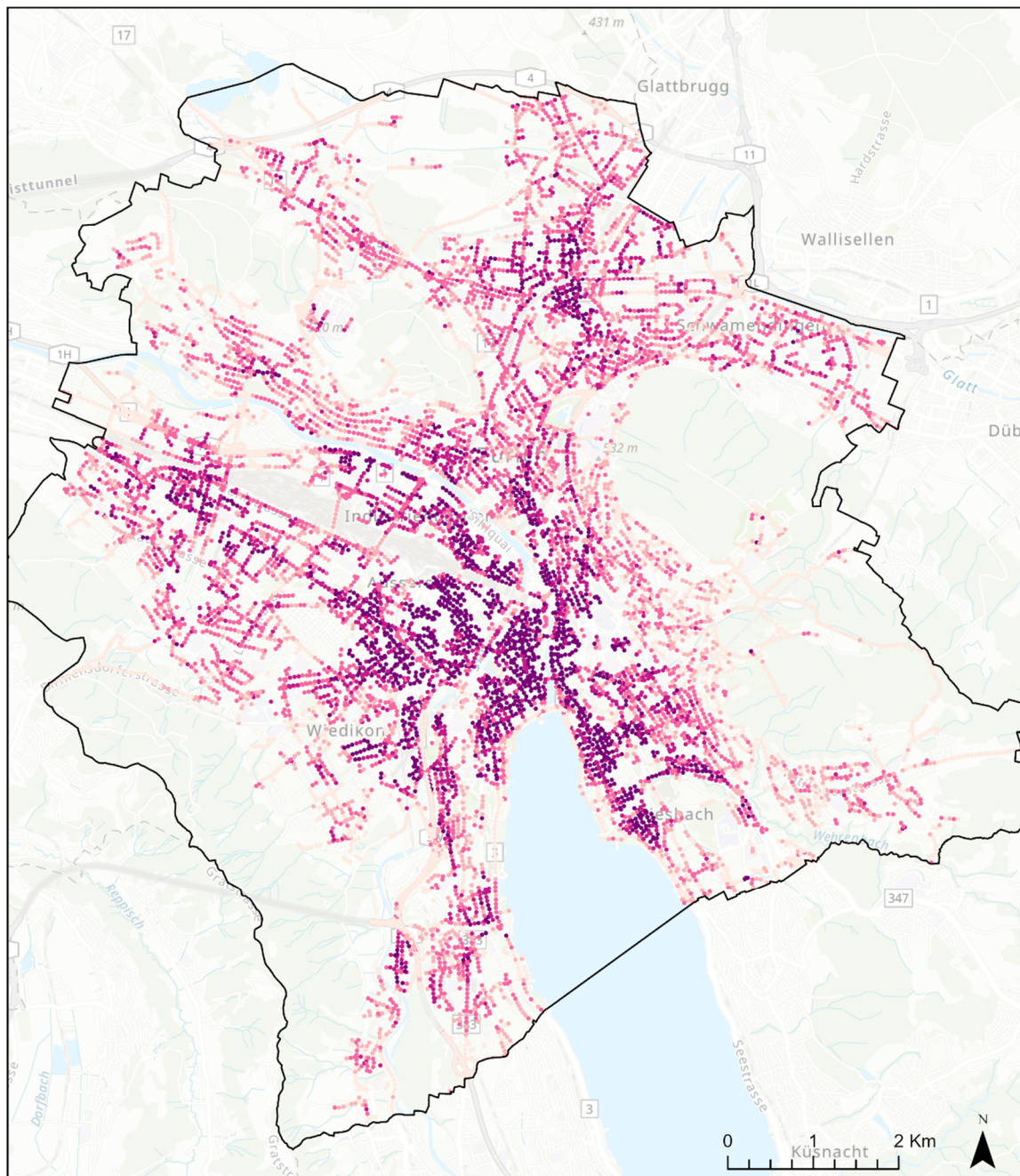
0,31 - 0,70

Spatial Reference

WGS 1984 UTM Zone 32N

Author

Toon Lambrecht, Lund
University (2025)



Spatial distribution of view factors for building-merged - Zurich

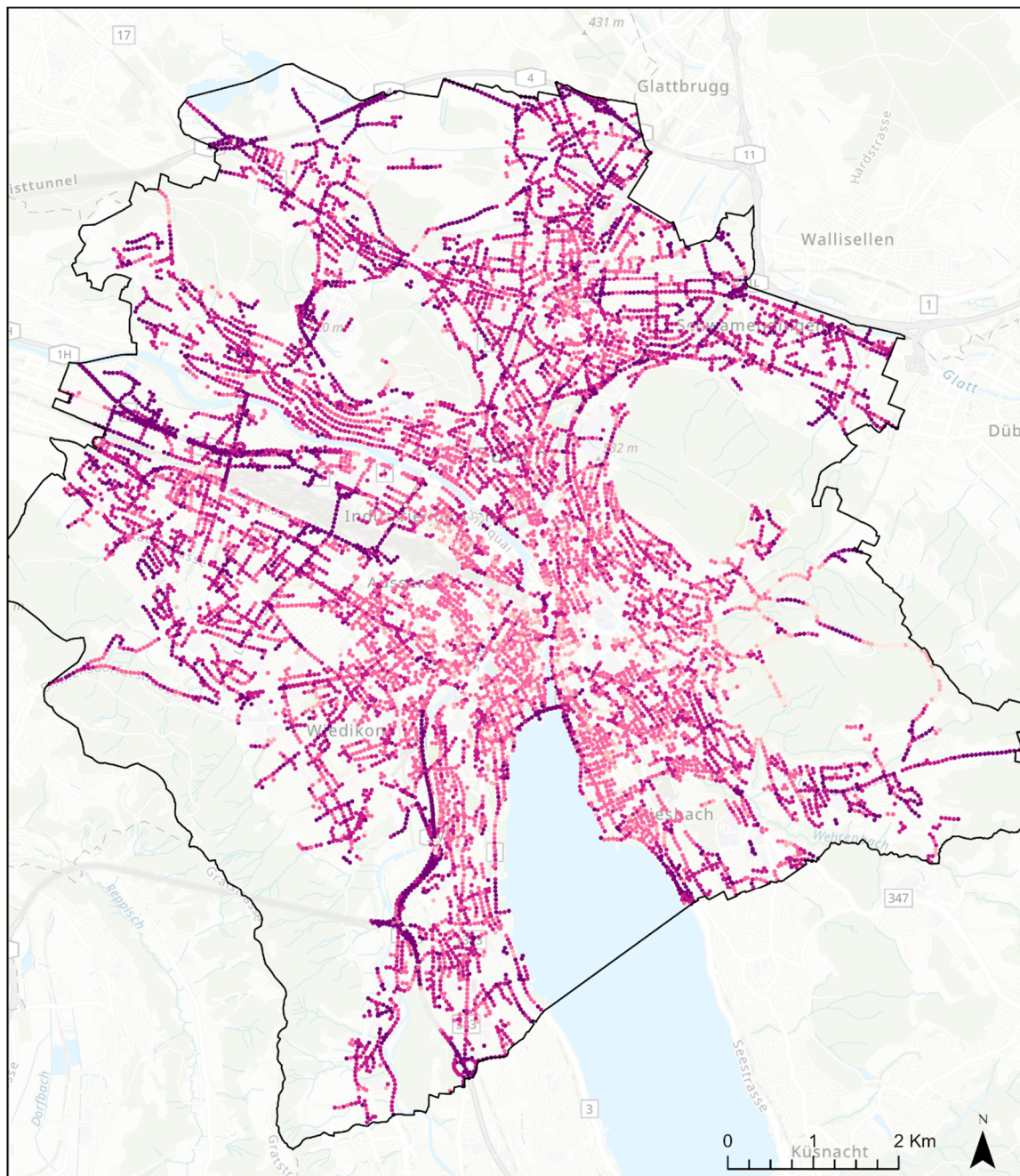
□ Zurich boundary

view factor

- 0,00 - 0,05
- 0,06 - 0,10
- 0,11 - 0,20
- 0,21 - 0,30
- 0,31 - 0,70

Spatial Reference
WGS 1984 UTM Zone 32N

Author
Toon Lambrecht, Lund
University (2025)



Spatial distribution of view factors for sky-merged - Zurich

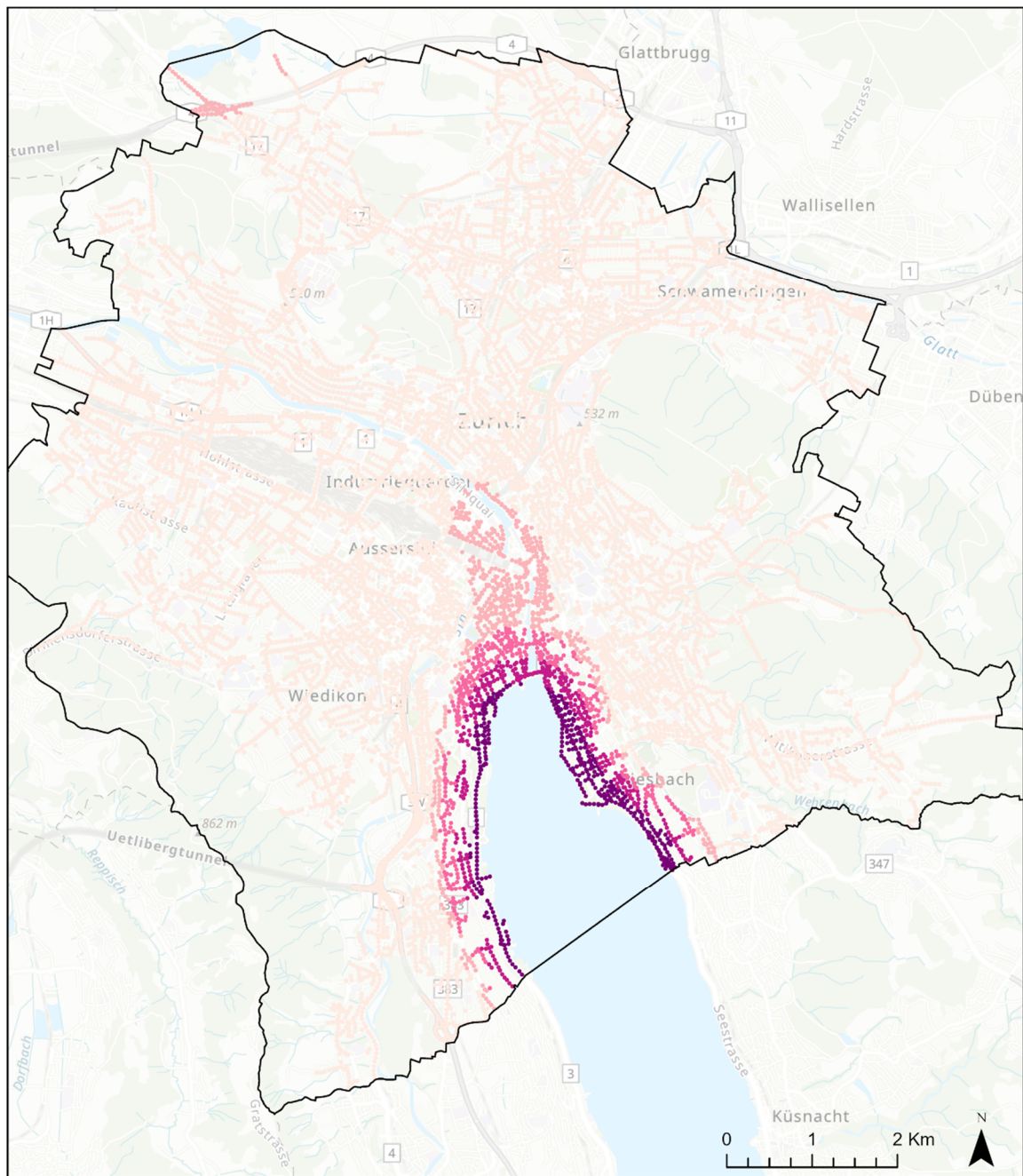
□ Zurich boundary

view factor

- 0,00 - 0,05
- 0,06 - 0,10
- 0,11 - 0,20
- 0,21 - 0,30
- 0,31 - 0,70

Spatial Reference
WGS 1984 UTM Zone 32N

Author
Toon Lambrecht, Lund
University (2025)



Spatial distribution for water area within 800m - Zurich

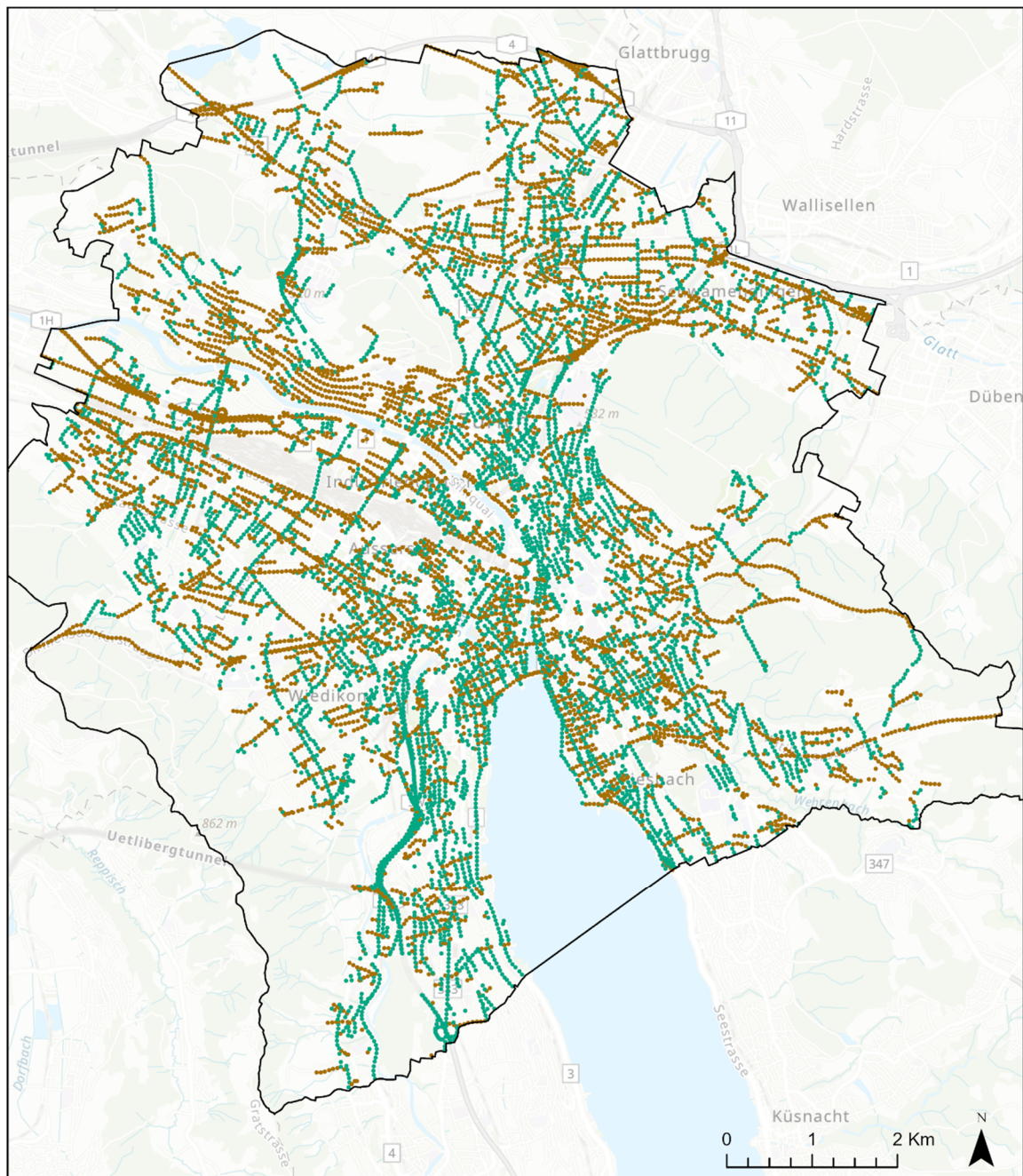
□ Zurich boundary

water area

- 0,00 - 0,05
- 0,06 - 0,10
- 0,11 - 0,20
- 0,21 - 0,30
- 0,31 - 0,70

Spatial Reference
WGS 1984 UTM Zone 32N

Author
Toon Lambrecht, Lund
University (2025)



Spatial distribution of street orientation- Zurich

□ Zurich boundary

orientation

● NS

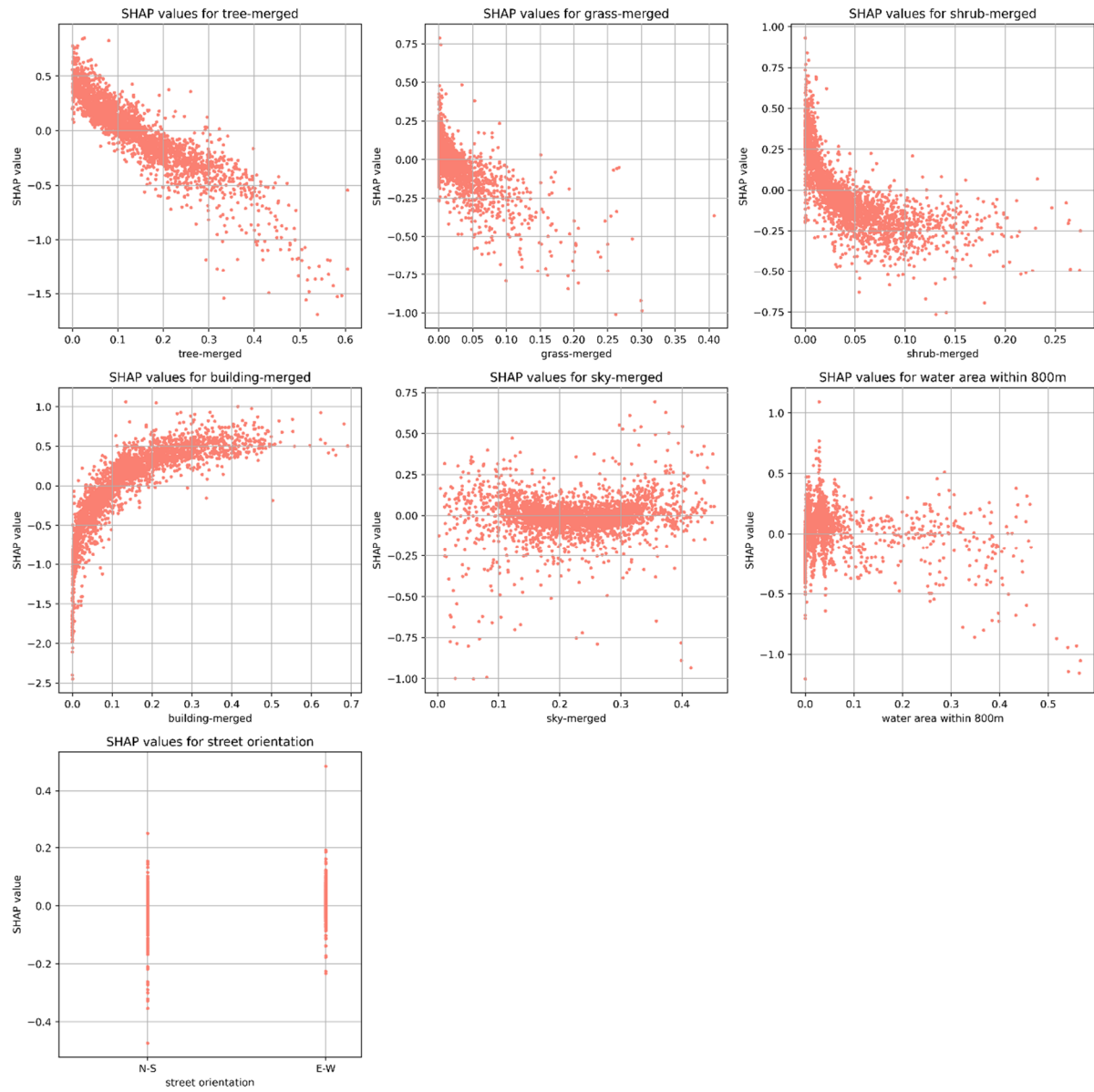
● EW

Spatial Reference
WGS 1984 UTM Zone 32N

Author
Toon Lambrecht, Lund
University (2025)

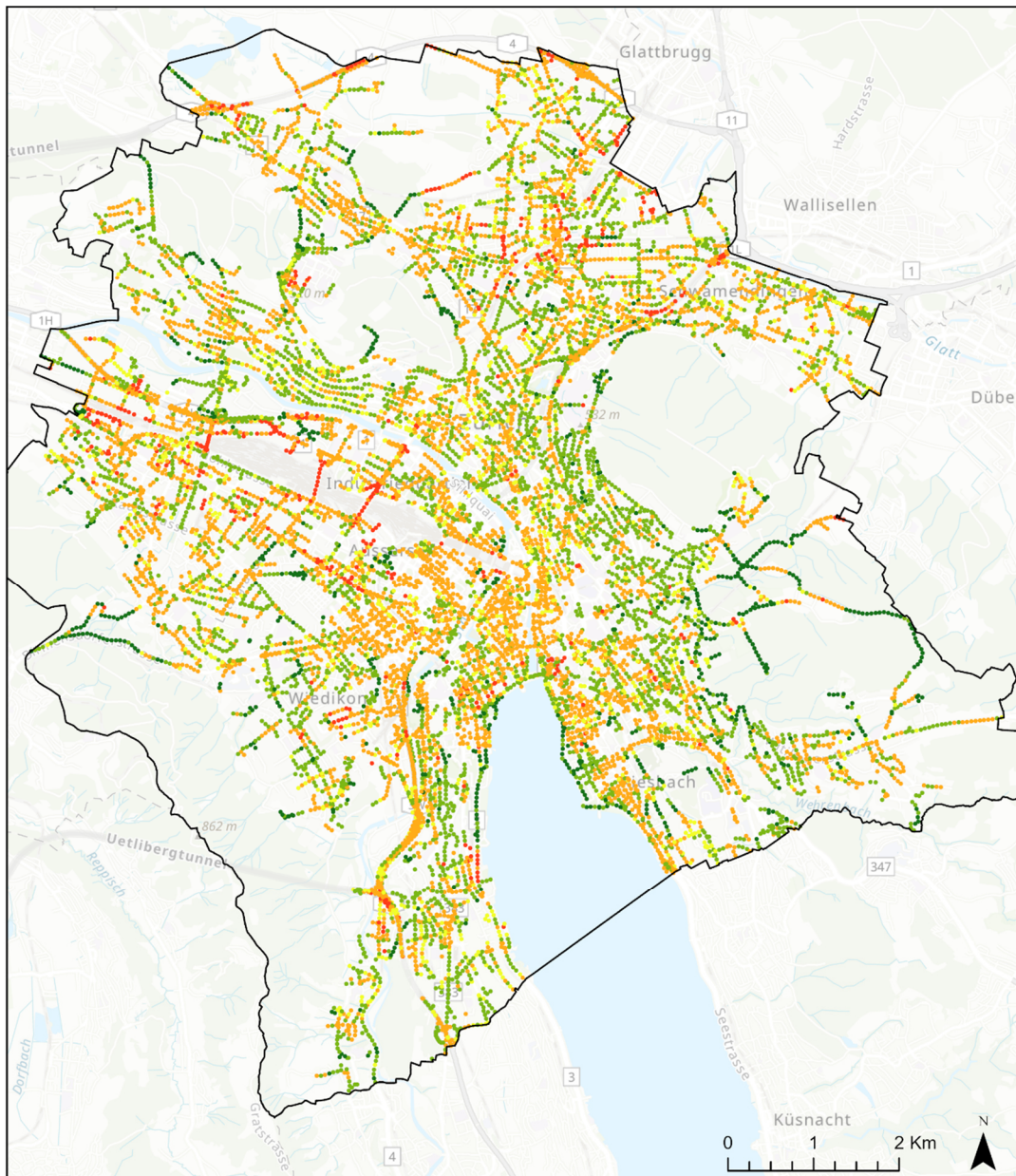
Appendix D: SHAP values from GWRf for test data

This appendix includes the scatters plots of the SHAP values for the test data according to their variable distribution. The SHAP values for the training data were presented in Figure 22.



Appendix E: Individual maps of SHAP values

This appendix includes the individual maps that are shown together in Figure 23. They were individually included to aid interpretation. Every map shows the contribution of the respective variable to LST in °C.



Spatial distribution of SHAP values for tree-merged - Zurich

□ Zurich boundary

SHAP Value (°C)

● -1,99 - -0,50

● -0,49 - -0,05

● -0,04 - 0,05

● 0,06 - 0,50

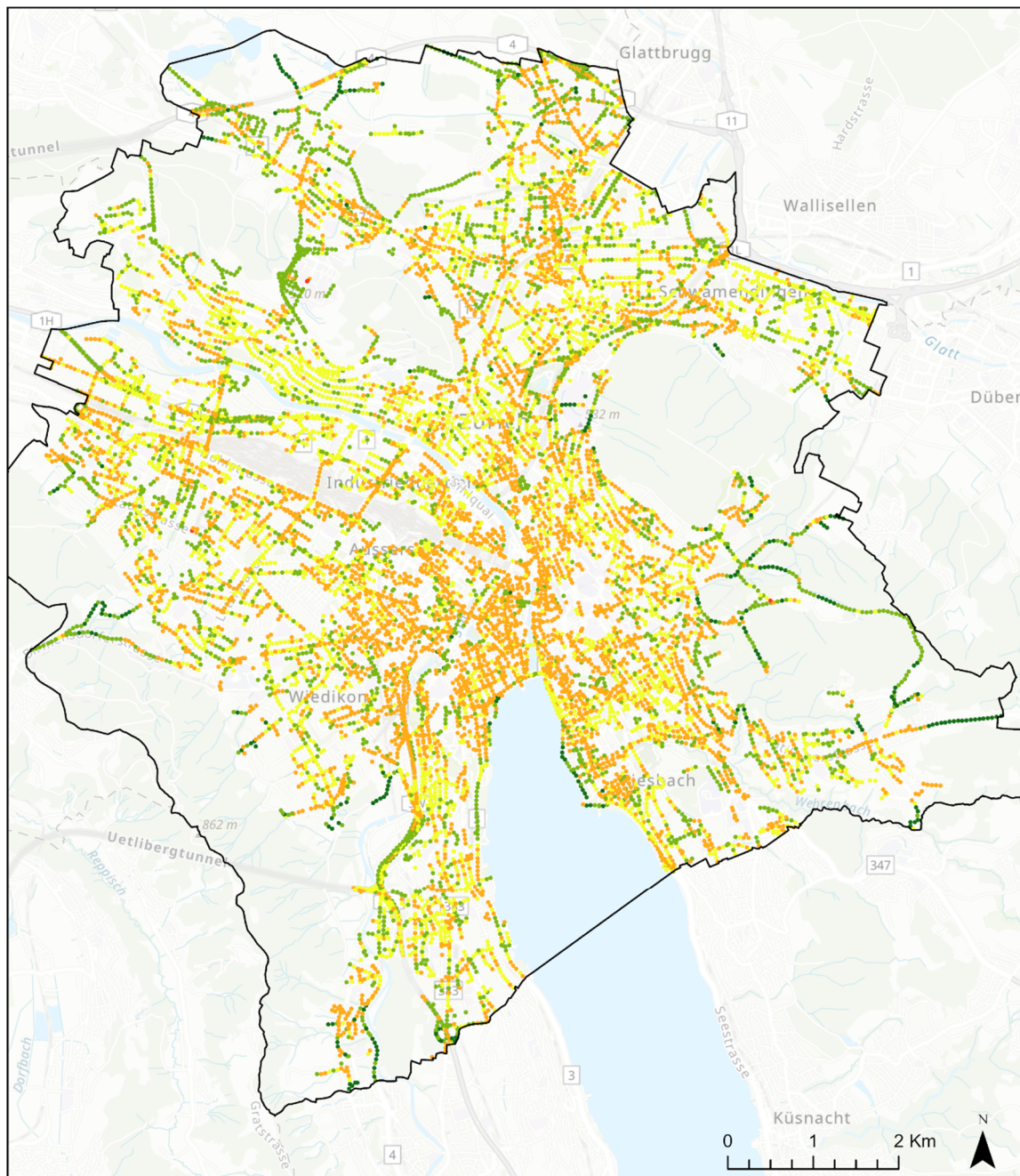
● 0,51 - 0,91

Spatial Reference

WGS 1984 UTM Zone 32N

Author

Toon Lambrecht, Lund
University (2025)



Spatial distribution of SHAP values for grass-merged - Zurich

□ Zurich boundary

SHAP Value (°C)

● -1,02 - -0,50

● -0,49 - -0,05

● -0,04 - 0,05

● 0,06 - 0,50

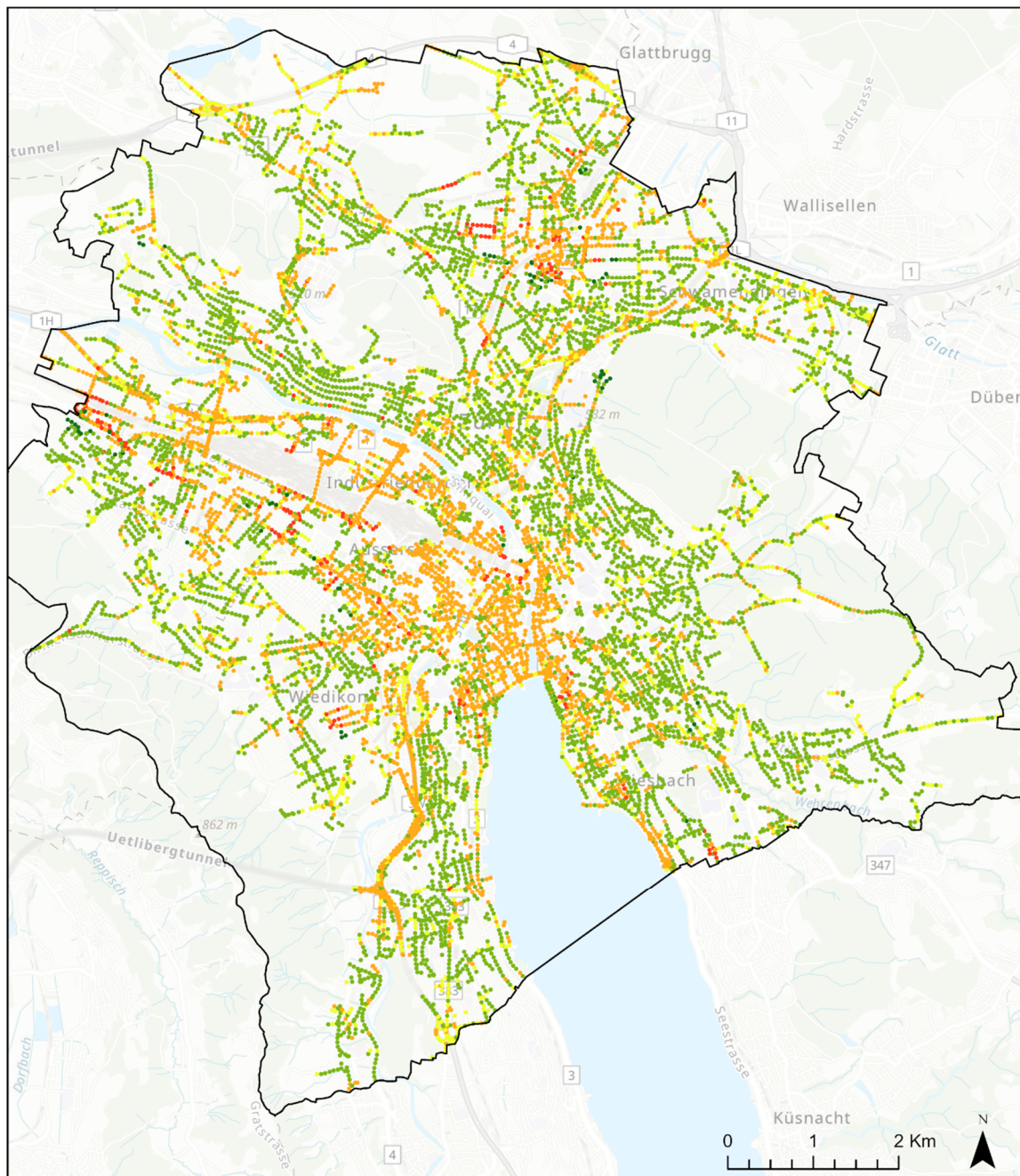
● 0,51 - 0,74

Spatial Reference

WGS 1984 UTM Zone 32N

Author

Toon Lambrecht, Lund University (2025)



Spatial distribution of SHAP values for shrub-merged - Zurich

□ Zurich boundary

SHAP Value (°C)

● -0,79 - -0,50

● -0,49 - -0,05

● -0,04 - 0,05

● 0,06 - 0,50

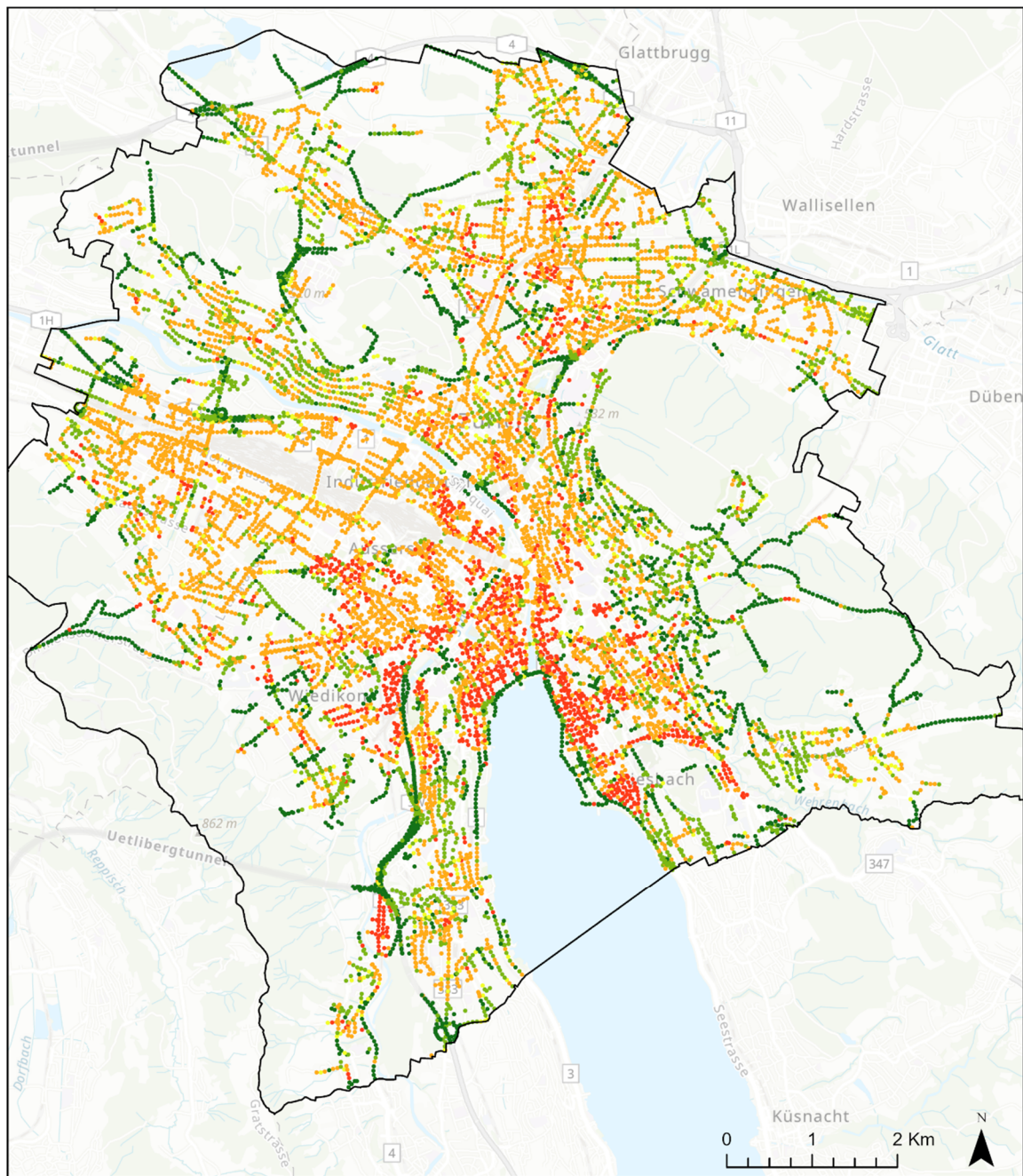
● 0,51 - 1,11

Spatial Reference

WGS 1984 UTM Zone 32N

Author

Toon Lambrecht, Lund
University (2025)



Spatial distribution of SHAP values for building-merged - Zurich

□ Zurich boundary

SHAP Value (°C)

● -2,45 - -0,50

● -0,49 - -0,05

● -0,04 - 0,05

● 0,06 - 0,50

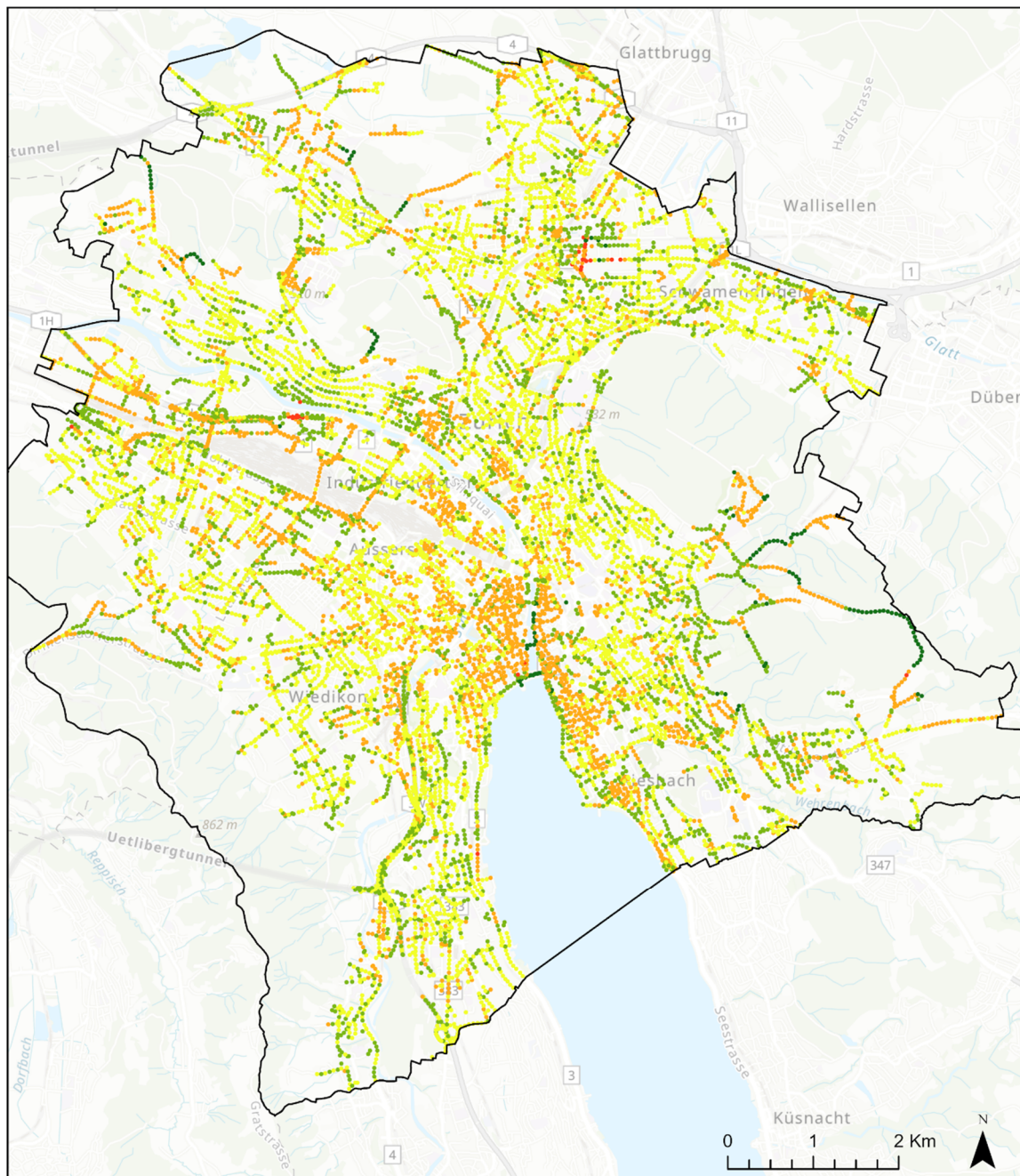
● 0,51 - 1,23

Spatial Reference

WGS 1984 UTM Zone 32N

Author

Toon Lambrecht, Lund
University (2025)



Spatial distribution of SHAP values for sky-merged - Zurich

□ Zurich boundary

SHAP Value (°C)

● -1,34 - -0,50

● -0,49 - -0,05

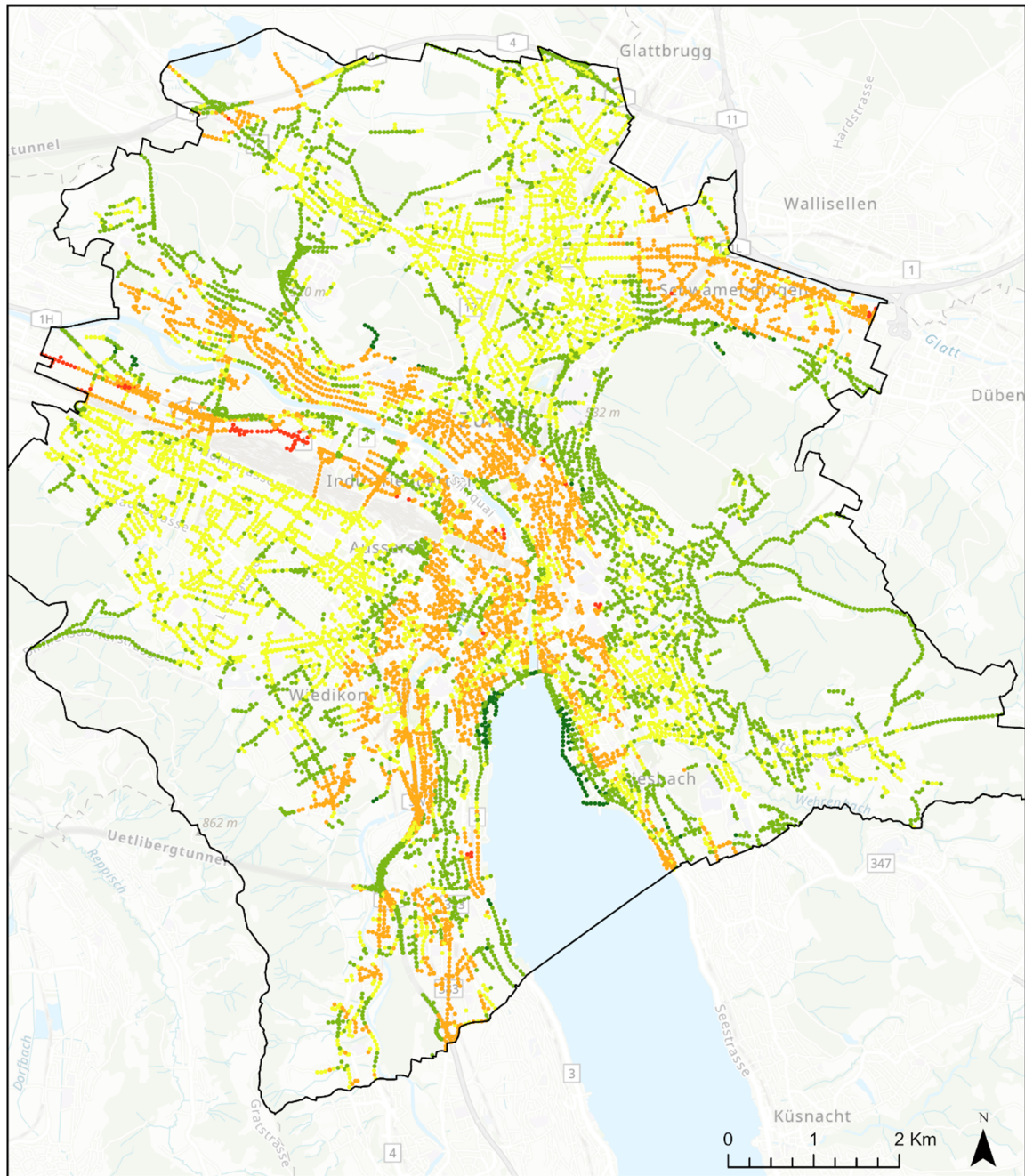
● -0,04 - 0,05

● 0,06 - 0,50

● 0,51 - 0,69

Spatial Reference
WGS 1984 UTM Zone 32N

Author
Toon Lambrecht, Lund
University (2025)



Spatial distribution for SHAP values of water area within 800m - Zurich

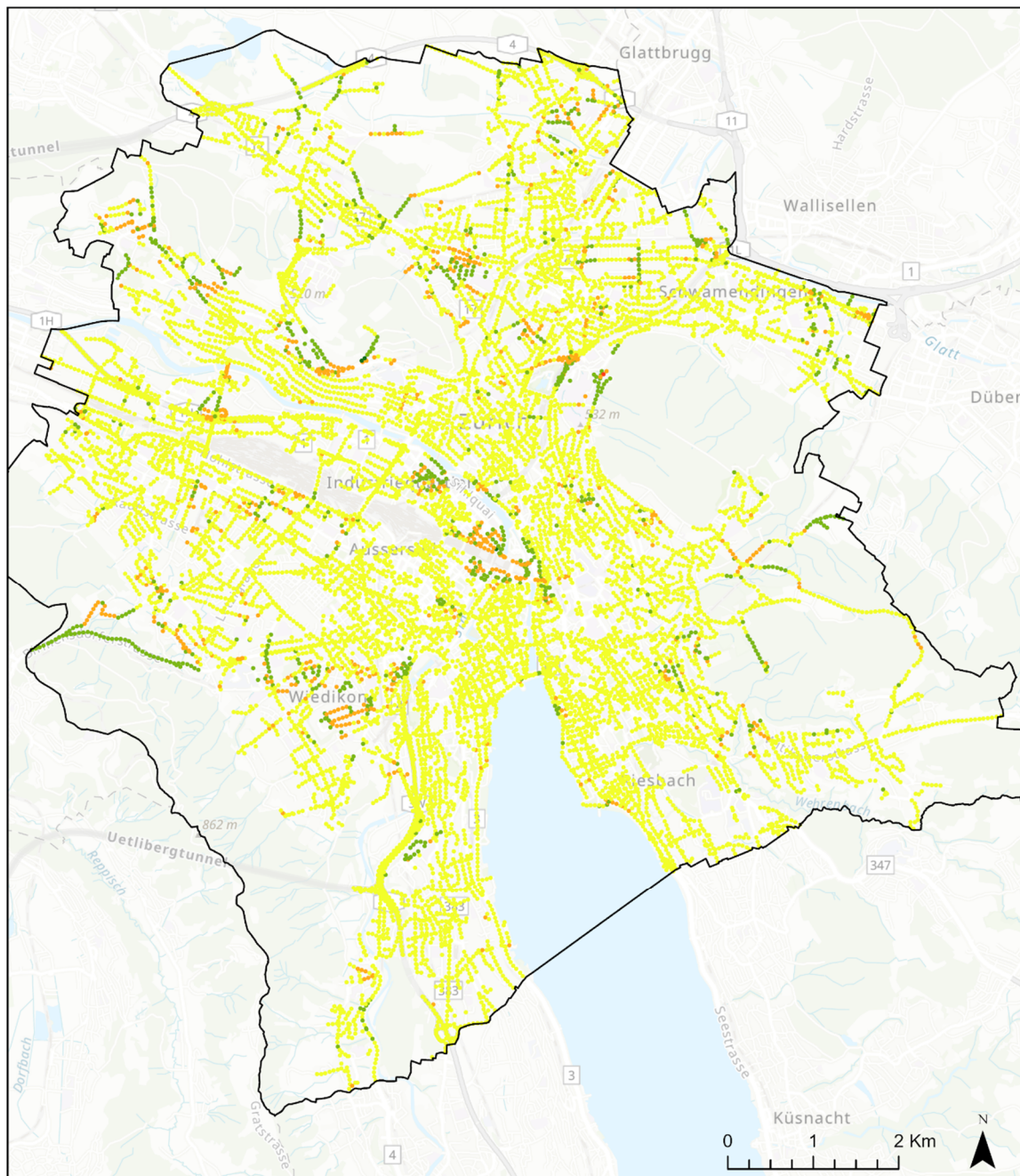
□ Zurich boundary

SHAP Value (°C)

- -1,21 - -0,50
- -0,49 - -0,05
- -0,04 - 0,05
- 0,06 - 0,50
- 0,51 - 1,11

Spatial Reference
WGS 1984 UTM Zone 32N

Author
Toon Lambrecht, Lund
University (2025)



Spatial distribution of SHAP values for orientation- Zurich

□ Zurich boundary

SHAP Value (°C)

- -0,47 - -0,47
- -0,46 - -0,05
- -0,04 - 0,05
- 0,06 - 0,50
- 0,51 - 0,54

Spatial Reference
WGS 1984 UTM Zone 32N

Author
Toon Lambrecht, Lund
University (2025)