Remote Sensing of Environment 287 (2023) 113454

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Predicting the liveability of Dutch cities with aerial images and semantic intermediate concepts

Alex Levering [a,*], Diego Marcos [b], Jasper van Vliet [c], Devis Tuia [d]

[a] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, The Netherlands
[b] Inria, University of Montpellier, France
[c] Institute for Environmental Studies, Vrije Universiteit Amsterdam, The Netherlands
[d] Environmental Computational Science and Earth Observation Laboratory, EPFL, Switzerland

ARTICLE INFO

Edited by Marie Weiss

Dataset link: https://github.com/ahlevering/liveability-rs

Keywords:
Liveability
Deep learning
Aerial imagery
Urban studies

ABSTRACT

In order to provide urban residents with suitable living conditions, it is essential to keep track of the liveability of neighbourhoods. This is traditionally done through surveys and by predictive modelling. However, surveying on a large scale is expensive and hard to repeat. Recent research has shown that deep learning models trained on remote sensing images may be used to predict liveability. In this paper we study how well a model can predict liveability from aerial images by first predicting a set of intermediate domain scores. Our results suggest that our semantic bottleneck model performs equally well to a model that is trained only to predict liveability. Secondly, our model extrapolates well to unseen regions (R^2 between 0.45 and 0.75, Kendall's τ between 0.39 and 0.57), even to regions with an urban developmental context that is different from areas seen during training. Our results also suggest that domains which are directly visible within the aerial image patches (physical environment, buildings) are easier to generalize than domains which can only be predicted through proxies (population, safety, amenities). We also test our model's perception of different neighbourhood typologies, from which we conclude that our model is able to predict the liveability of neighbourhood typologies though with a varying accuracy. Overall, our results suggest that remote sensing can be used to extrapolate liveability surveys and their related domains to new and unseen regions within the same cultural and policy context.

1. Introduction

The living standards of a neighbourhood may have a significant effect on the health of residents. Residents of destitute neighbourhoods are prone to several health risks, such as increased morbidity rates (Barber et al., 2016), mortality rates (Haan et al., 1987), and worse dietary and physical activity patterns (Thompson and Kent, 2014). Similar patterns are observed for housing, where lower-quality housing also results in worse mental well-being (Evans, 2003). As such, it is important to monitor the wellbeing of a neighbourhood for the benefit of urban residents. For this purpose, researchers have studied how factors relate human wellbeing to their living environments using the *liveability* framework. The liveability of a society can be understood as *"the degree to which its provisions and requirements fit with the needs and capacities of its members"* (Veenhoven et al., 1993). In the context of living environments, examples of the needs and capacities required may be housing that is of adequate size and quality for its residents, provision for adequate travel to work, and sufficient green space in the neighbourhood. Research has since advanced the theoretical underpinnings of liveability research. van Kamp et al. (2003) argue that a conceptual framework of liveability would *"allow for a more theory-based choice of indicators, and for the development of tools to evaluate multidimensional aspects of urban environmental quality"*. The Leefbaarometer project (referred to as *LBM* hereafter) initiated by the Dutch government (Leidelmeijer et al., 2014) follows up on that suggestion. The LBM project was set up to survey the liveability of neighbourhoods across the Netherlands, and to subsequently model the liveability using variables that can be applied nation-wide, such as housing quality and greenspace proximity. In doing so, the authors assess which variables are relevant for liveability at a nation-wide scale. Linking such survey data to empirical and statistical data may improve our understanding of what makes cities liveable. However, a notable drawback to using manually collected data such as surveys is the difficulty in upscaling and repeating results.

Remote sensing methods have long been used to extract intermediate variables for liveability prediction, such as the prediction of urban

[*] Corresponding author.
E-mail address: alex.levering@wur.nl (A. Levering).

https://doi.org/10.1016/j.rse.2023.113454
Received 29 September 2022; Received in revised form 2 December 2022; Accepted 10 January 2023
Available online 24 January 2023

greenery (Jensen et al., 2004; Li and Weng, 2007; Rahman et al., 2011), rather than the prediction of liveability directly from imagery. Studies attempting to recognize the qualities of cities have considered various intermediate variables, such as urban morphology (Taubenböck et al., 2012; Rodriguez Lopez et al., 2017; Tian et al., 2022), local climactic conditions (Bechtel et al., 2015; Qiu et al., 2019; Liu and Shi, 2020), and urban land use (Srivastava et al., 2019; Rosier et al., 2022). Recent advances in machine and deep learning have enabled research which predicts liveability variables directly from overhead imagery. Remote sensing models have the benefit of high scalability and better monitoring in regions with poor data availability (Kuffer et al., 2020, p. 18). In regions with greater data availability, much research has gone into hedonic housing pricing as a means of predicting the attractiveness of neighbourhoods. Hedonic housing pricing attempts to capture the value of a property based on its intrinsic value, as well as external factors affecting it. The main value of remote sensing for hedonic pricing is the inclusion of contextual information about the immediate and larger area of surroundings (Bency et al., 2017, p.5). Yao et al. (2018), for example, fuse remote sensing imagery with social media data to predict housing prices in Shenzhen, China, with highly accurate results.

Recent studies have attempted to directly predict variables relating to liveability in countries with high data availability. Arribas-Bel and colleagues trained machine learning models to recognize living environment deprivation from high-resolution aerial images over the city of Liverpool in the United kingdom (Arribas-Bel et al., 2017). Singleton et al. (2022) use an autoencoder model to extract features describing Sentinel-2 satellite image tiles of neighbourhoods across the UK. These features were clustered to form neighbourhood typologies, and subsequently related to urban deprivation data. However, the clustered neighbourhood representations proved insufficient to explain urban deprivation. Suel and colleagues study income, overcrowding, and environmental deprivation using a multimodal approach, using both Google Street View and 3 m resolution Planet satellite images over the Greater London region (Suel et al., 2021). Their findings confirm that high-resolution aerial images on their own can approximate the trend of urban deprivation at the neighbourhood level. Scepanovic et al. (2021) use Sentinel-2 image tiles to predict the vitality (presence of people throughout the day) of Italian cities at the district level through several experiments. The authors predict 6 physical descriptors of urban form relating to land use and block size from Sentinel-2 image patches across Italian districts and infer their usefulness for predicting vitality. This first experiment showed limited accuracy, most likely due to the resolution of the Sentinel-2 image tiles. In their second experiment, the authors predict urban vitality (as measured by mobile internet usage) directly from Sentinel-2 image features, and the capacity of models to generalize between cities. Their results indicate that generalization of urban vitality is possible, but generalizing their model to Rome resulted in a notable decrease in accuracy, as it is historically, culturally, and naturally distinct from the other cities within their dataset. Huang and Liu (2022) use a deterministic approach to model liveability of 101,630 communities in China in 42 major cities, guided by expert decisions. A total of 27 liveability factors are extracted using high-resolution satellite imagery and subsequently weighted according to expert opinions. Their work presents the first large-scale assessment of the liveability of urban communities in China.

Previous work has attempted to study remotely-sensed liveability by observing a limited number of components relating to liveability at a time, and without taking into account surveyed resident opinions. In doing so, they have confirmed that individual liveability factors such as income, environmental deprivation, and block size can suitably be predicted through optical remote sensing. Yet, it is unclear to what extent different domains relating to liveability can be predicted from remote sensing imagery. Therefore, in this paper, we study how well different domains of liveability may be predicted from high-resolution aerial imagery on a neighbourhood scale. We set out to determine the suitability of remote sensing for interpolating and extrapolating large-scale inventories of liveability. Moreover, we explore how a model with a semantic intermediate layer compares to a model which only predicts liveability. Specifically, we compare how the liveability prediction as a linear combination of domain scores compares against a direct prediction of liveability. Lastly, we evaluate how well liveability domains can adapt to unseen geographical contexts, as well as building typologies. We formulate and address two research questions for our research:

1. How well can we predict different domains of liveability?
2. How well does a bottleneck model predict compared to an unconstrained model?

The remainder of the paper is as follows. In Section 2 we present the dataset used in our study and our model architecture. In Section 3 we present the metrics and maps for our experiments. Lastly, in Section 4 we reflect on our results and their relevance for liveability monitoring.

2. Methods

We are interested in training a deep learning model to predict liveability on a neighbourhood scale by first predicting domain-specific liveability contribution scores as a set of interpretable semantic intermediate concepts. For this purpose we use a semantic bottleneck model (Marcos et al., 2021; Koh et al., 2020), which uses an intermediate linear layer with semantic concepts which are then used to predict a final objective. For this purpose we need a dataset of overhead aerial images, neighbourhood-scale labels of liveability, and a deep learning model architecture which can first predict individual domain scores, and then regress the overall liveability score through the domain specific scores. We discuss these requirements in order.

2.1. *Dataset design*

To train our model we require a labelled dataset of liveability scores and overhead aerial imagery (Fig. 1). Additionally, we make use a series of domain scores, which decompose the liveability score into a series of explainable aspects. To build this dataset we use nationally available data sources in the Netherlands. Specifically, we consider 13 built-up areas of varying sizes, ranging from village (Beesel) to metropolis (Amsterdam). Selected built-up areas are listed in Table 1.

2.1.1. *Liveability reference data*

The reference data for liveability used in our research is made available by the Leefbaarometer (*LBM*) project (Leidelmeijer et al., 2014), an ongoing liveability monitoring project initiated by the Dutch government. For this purpose, the authors collected a dataset with over 100 variables for use in regression models to predict liveability. These variables are available for all neighbourhoods in the Netherlands at the scale of an individual street. Where applicable, variables are summed over a radius of 200 m around each neigbhourhood to reduce the occurrence of outlying neighbourhood with few respondents in the dataset. The input variables can be designated to five domains. The following broad groups of variables are considered for each domain:

- **Population**: Welfare factors, age groups, residuals for family composition and ethnic composition after controlling for income
- **Physical Environment**: Green/gray area descriptors, proximity to water/green areas, proximity to nuisances (e.g. trains/roads)
- **Safety**: Number of occurrences for several broad crime categories
- **Amenities**: Amenities within 1–20 km distance, e.g. cafes, hospitals, schools
- **Buildings**: Building age groups per 10 years after 1900, ownership status, simple typology descriptors e.g. pre/post-war

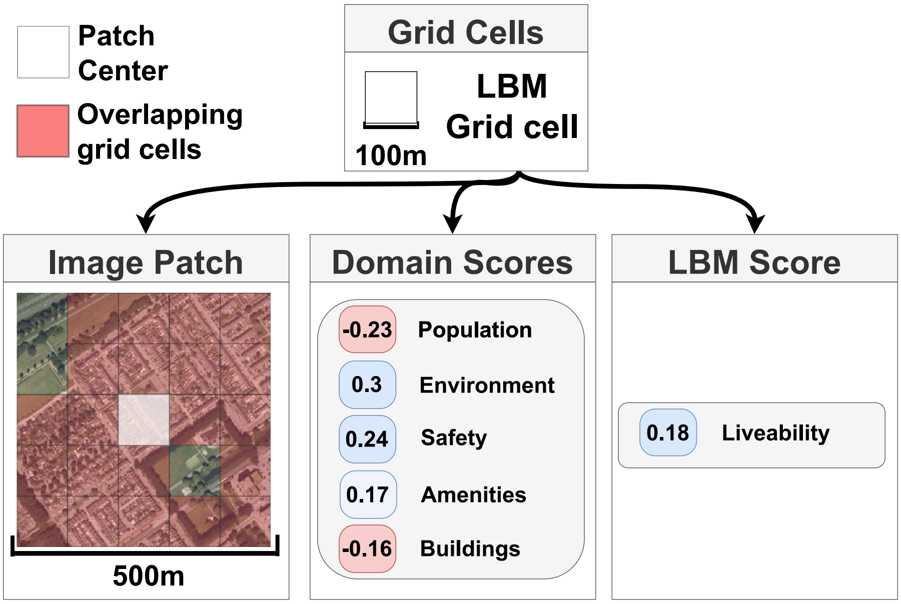

Fig. 1. Workflow for generating our reference data. From the LBM dataset we extract the domain scores and the final liveability score. The domain scores are a decomposition of the liveability score which reflect how each domain contributes to the overall liveability of a grid cell. For our image patches we use the grid cell as the center for a 500 by 500 pixels patch at 1 m resolution. The 400 by 400 m overlap with other grid cells ensures that the patch size is equal to the spatial sum operation which was performed for the original variables of the LBM dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For a complete description of all 100 variables used by the LBM project we refer the reader to table 7.1 of the documentation (Leidelmeijer et al., 2014, p.91). In the discussion we elaborate on the use of stigmatizing variables for the population domain score and the problems arising from it.

These 100 variables belonging to five domains are then used as the input for two linear regression models. The first model regresses surveyed resident liveability opinions. Respondents were asked three questions about their satisfaction with their living situation and asked to answer on a scale of 1–5 for each question, where 5 is "*most satisfied*". The average of these three questions is used as the response variable for the first regressor, after correcting for the age of residents. The second model uses a hedonic pricing approach to estimate housing prices for a neighbourhood derived from nationally-available property value estimates. From these two linear regression models, each neighbourhood is assigned the averaged z-score of these models as the single overall liveability score, shown on the right side of Fig. 1. Hereafter, we will refer to this averaged z-score simply as the *liveability score*. By grouping the 100 variables into five domains and by averaging their coefficients, the contribution of each domain to the overall change in z-score can also be computed for each domain. We refer to these grouped scores as *domain scores*. The five domain scores are fundamentally different in nature. Some domain scores can be observed directly from aerial images. This group consists of the *buildings* and the *physical environment* domain scores. We refer to these scores as *direct* scores. The other three domain scores cannot be observed from aerial images, but should instead be predicted by proxy correlations. For instance, for the *Population* domain score, the model could learn that large single-family houses generally have a more affluent population, thereby learning a correlation as a proxy for the prediction of the domain score. We refer to these domain scores (*Population*, *Safety*, and *Amenities*) as *indirect scores*.

The veracity of the outcomes of the LBM project was verified through interactions with policy makers. For all of the 13 built-up areas considered in this research the results truthfully reflected the general liveability trends (Leidelmeijer et al., 2014, p.100). The liveability score and the domain scores are re-predicted bi-yearly from 2014 onwards. For privacy reasons, the dataset could not be made available at the street level. Instead, all variables and scores are made available at the resolution of 100 meters through a gridded dataset. We use the grid cells made available in this research as the basis for our dataset, for both their spatial extent and as reference data.

2.1.2. *Neighbourhood liveability patches*

We use the gridded dataset provided by the LBM project as the starting point for our dataset of neighbourhood liveability patches. We use the liveability scores made available for the year 2016. We do so firstly because it is the closest year to which there is a nationally-available aerial image (2016). In total we use 51,781 grid cells from the dataset over the 13 built-up areas within our dataset. The samples used from each built-up area are shown in Table 1. We use the five domain scores and the overall liveability scores (middle and right columns of Fig. 1 respectively) as the liveability labels of our patches.

For the overhead aerial imagery we use images from the national composite aerial image from 2017, made available by the Dutch government (PDOK, 2017). The original composite image is available at 0.1 m resolution with four bands (red, green, blue, near-infrared (NIR)) and is entirely cloud-free. We do not perform additional pre-processing steps such as geometric correction, as this has already been done by the data provider. We downsample the pixel size to 1 m.

Beyond determining how well liveability can be predicted, we are interested in monitoring it over multiple timesteps. However, high-resolution imagery available for past years does not have NIR information. To ensure the compatibility of our analyses with historical aerial image data in The Netherlands, we exclude the NIR band from our main analyses. In future work we will explore the feasibility of time series mapping for liveability. However, we study the effect of adding the NIR band in our liveability prediction model in the results and discussion section, where we report the numerical results for a model trained on all four bands.

As some LBM variables are summed over a radius of 200 m around the grid cells, the square patch size should cover at least 500 by 500 meters such that it approximates the extent of the LBM grid cell centers. As such, we extract patches of 500 by 500 pixels centered on each grid cell center. As a result of the image patch being larger than the 100 by 100 m LBM grid cells, there is an overlap with the 24 bordering neighbouring aerial image patches.

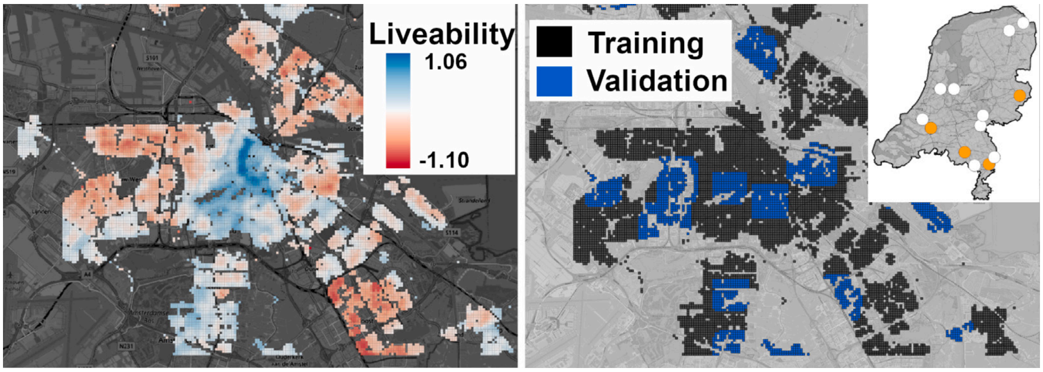

Fig. 2. Left: Liveability scores over Amsterdam, ranging from −1.10 (lowest, red) to 1.06 (highest, blue). **Right**: Example of data splits. Grid cells marked with dark grey are used during training. Blue grid cells are used for validation. In the top-right we show built-up areas which are considered. Areas marked with white points are used for training/validation, while areas marked with orange points are used only during inference. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Samples per split and municipal population census numbers for each built-up area. Population data is derived from the Dutch statistics agency (CBS, 2016).

Built-up area	Training	Validation	Testing	Population (2016)
Almere	1,856	1,206	–	198,145
Amsterdam	7,116	2,609	–	833,624
Arnhem	3,713	722	–	153,818
Beesel	–	–	388	13,388
Dordrecht	–	–	3,548	118,801
Eemsdelta	607	238	–	47,080
Eindhoven	–	–	6,490	224,755
Groningen	2155	718	–	200,952
Hengelo	–	–	3,034	81,075
Nijmegen	3,071	1,068	–	172,064
Rotterdam	8,439	1,823	–	629,606
Venlo	1,074	664	–	100,371
Weert	1,008	234	–	49,100
Total	29,039	9,282	13,460	–

2.1.3. Data splitting

We use data from 9 built-up areas for training and validation. Within each area we create square blocks of patches for validation, and we assign the rest to the training set. Through the overlap with neighbouring patches, some of the validation set is seen during training. However, this was not found to result in issues with generalization during testing. We use the remaining 4 built-up area as an independent test set. The 4 cities were chosen for their geographic diversity and their size. Dordrecht is proximate to Rotterdam and it is part of the *Randstad* area, which is the largest conurbation of the Netherlands. As Amsterdam and Rotterdam are part of the training dataset, Dordrecht is therefore the most similar city in the test set. Eindhoven and Hengelo are both cities which follow a different development pattern compared to those in our training split. Both cities began to develop significantly as a result of industrialization, which makes them developmentally distinct from the cities in our training split. This difference in developmental context allows us to study how well our model adapts to unseen developmental layouts. Lastly, Beesel is a small village along the German border, which tests the model's ability to transfer to smaller settlements (as Beesel is the only village in the training dataset), and to remote regions.

We show an example of our training/validation set stratification for the municipality of Amsterdam in Fig. 2. We show the number of samples per split in Table 1.

2.2. Bottleneck CNN for liveability prediction

In this section we present the interpretable bottleneck model used to predict liveability from overhead aerial images. We use a two-step approach to predicting the overall liveability score of an area. To obtain a transparent and interpretable prediction of liveability which is concordant with the design of the LBM scores, we use a semantic bottleneck design (Marcos et al., 2021; Levering et al., 2020). A semantic bottleneck forces the prediction of a final layer to be interpretable by first predicting a set of semantic concepts, which are then linearly re-combined to predict the target variable. We chose this type of architecture because the LBM is by design a combination of the five domain scores. We can therefore use the semantic bottleneck to enforce the prediction of the liveability to be a linear combination of the domain scores, and this mimic the logic of the original LBM model. As such, our model is tasked with predicting concepts as a vector of domain-specific sub-scores, which we denote with $\mathbf{d}$. These domain scores are then used in a linear layer with a bias term to regress the predicted patch liveability score l. Our architecture is shown in Fig. 3.

Our model is first tasked with extracting relevant features for the prediction of liveability. The feature extractor takes the aerial image patch as input and produces a global feature vector $\mathbf{r}$. We use a standard convolutional neural network feature extractor for this purpose. Using this global feature vector $\mathbf{r}$ we then predict a liveability domain score for each of the $i \in \{1 \dots D\} \in \mathbb{N}$ domains being considered. These liveability domain scores describe the contribution of different domains to the overall liveability of a place in explainable aspects, such as *amenities* and *safety*. The domain scores correspond to the domain scores presented in the middle columns of Fig. 1. To predict the domain scores, we use a two-layer Multi-Layer Perceptron (MLP) to create each row of the feature matrix $\mathbf{C}$. The first linear layer re-combines the extracted features into a 250-dimensional vector which is activated by a ReLU non-linearity. Notice that this feature vector $\mathbf{C_i}$ represents a summary of the features as they are relevant for each domain, which we leverage when interpreting the model's propositions in Section 3.2.2. The second layer of each MLP uses the feature vector $\mathbf{C_i}$ to regress the domain-specific liveability sub-score, which are the scores in the middle column of Fig. 1. We then concatenate all of the liveability domain scores to form the domain score vector $\mathbf{d}$. From the liveability domain score vector $\mathbf{d}$ (plus a bias term) we then directly regress the overall liveability score $\hat{l}$. In doing so, we enforce that the overall scenicness is only predicted by the linear combination of domain scores, rather than spurious correlations which the model may pick up on from the aerial images. As the domain scores are predicted as an intermediate task in our model, we can assess their accuracy to determine how well liveability domain scores can be predicted (research question 1).

Our model is trained using a combination loss of the domain score losses and the liveability score loss. The domain score loss is given as the sum of the mean squared errors over all of the domain scores w.r.t. their reference score $\hat{d}_i$:

$$\mathcal{L}_{domain} = \sum_{i=1}^{D} (\mathbf{d}_i - \hat{\mathbf{d}}_i)^2 \tag{1}$$

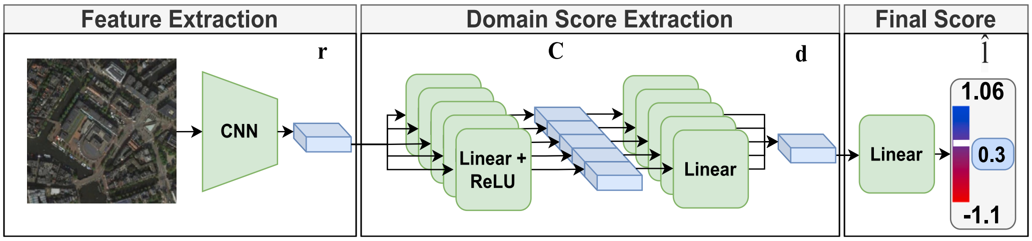

Fig. 3. Architecture of our model. Using a CNN we first extract a vector of features **r**. We then construct the rows of feature matrix **C**, where each row is a feature vector $\mathbf{C_i}$ that is specific to one domain score. Each feature vector is then used to compute a domain score d_i. Finally, the domain score vector **d** is used to compute the overall patch liveability score $\hat{l}$.

The loss of the liveability score is the mean squared error w.r.t. the reference score $\hat{l}$:

$$\mathcal{L}_{final} = (l - \hat{l})^2 \quad (2)$$

Finally, we combine both scores to create the overall loss to propagate:

$$\lambda\mathcal{L}_{domain} + \mathcal{L}_{final} \quad (3)$$

where λ is a weighting term set empirically to regulate the importance of the domain scores compared to the liveability score prediction.

2.3. Set-up

Our feature extraction model is a ResNet-50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) from which we remove the final fully-connected layer. Our model is trained on a single NVIDIA TitanX GPU with a batch size of 20. We optimize our models using the AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of $5e^{-5}$ and weight decay rate of $1e^{-4}$. We train our model for 15 epochs. To prevent that the model learns a set of features unrelated to the domain scores in feature matrix C, we set the weighting term λ of Eq. (3) to 100 such that the model favours the correct prediction of the intermediate task over the predictions of the final linear layer.

We assess the quality of our results with three metrics for each of the five domain scores of the LBM dataset, as well as for the overall liveability score. Firstly, we calculate the root mean squared error as a measure of error for all scores. Secondly, we calculate the coefficient of determination R^2 to determine the quality of the fit for each score. Lastly we compute Kendall's τ (Kendall, 1938) which measures the ranking of neighbourhood patches. This is possible because the liveability scores in our dataset may also be interpreted as ordinal variables, in the sense that the quality of each neighbourhood can be compared to every other neighbourhood, which represents a ranking. Kendall's τ ranges between a perfectly inverse correlation given as -1 to a perfect correlation given as 1.

In order to assess how well a bottleneck model performs compared to an unconstrained model (research question 2) we also train an unmodified ResNet-50 model. This unmodified model is tasked with predicting the overall liveability score without the semantic bottleneck and serves as a baseline against which the bottleneck model is compared.

Lastly, we train a model on the aerial image patches with the NIR band included to assess how this effects liveability prediction. We use the same hyperparameter selection and we initialize the network using pre-trained ImageNet weights (Deng et al., 2009). We use the weights from the red band for the NIR input channel.

2.4. Feature vector analyses

In order to understand our model's perceptions of the 5 domain scores and the liveability score, we design a series of feature vector analysis experiments which help to explain how the model observes the different domains of liveability. We use t-SNE embeddings and neighbourhood typology data to further typicate how our model observes urban spaces.

2.4.1. t-SNE embeddings

We assess the model's visual perception of the different neighbourhood typologies. To do so, we perform t-SNE dimensionality reduction (Maaten and Hinton, 2008) to visualize the latent space of the feature vectors of our model. t-SNE iteratively projects the high-dimensional space into a lower number of dimensions while preserving their neighbourhood structure in the original high-dimensional space. By doing so we can reduce the feature vectors to just two dimensions while respecting the non-linear relationships learned by the model in the original high-dimensional space. This allows us to visualize which patches are considered visually similar by the model. We perform t-SNE dimensionality reduction on the *buildings* row of the domain feature matrix, which is $C_{building}$, and the global feature vector r. We use a perplexity (balance between global and local patterns) of 100, a learning rate of 500, early exaggeration (tendency for clusters to become compact) of 150, and we run our model for 1,000 iterations. We consider all patches in the dataset, rather than just the test set patches in order to analyse data structures across the 13 built-up areas. We can then overlay the neighbourhood typologies of each patch for each point in the reprojected 2-dimensional space, allowing us to infer the visual homogeneity of neighbourhood typologies for that particular score.

2.5. Neighbourhood typologies

The Netherlands has an long history of spatial planning and zoning, which has been extensively described and documented in official policy and literature (Ministry of Infrastructure and the Environment, 2012). Over the years there have been many different planning philosophies intended to address the housing needs at the time. The LBM project did not explicitly take into account the neighbourhood planning styles, but rather used decade-spanning building age groups. As such, the neighbourhood typologies can be considered a more complete description of the neighbourhood style compared to the age brackets of the LBM. We perform two experiments using the neighbourhood typologies. Firstly, we assess how well our model is able to perceive the liveability of neighbourhood typologies through scatterplots which compare the predicted liveability to its reference value for each patch with a significant amount of a given typology. Secondly, as part of our feature vector analyses, we can assess how our model perceives the homogeneity of different typology styles, as well as the links between certain planning styles as defined by Dutch planners. It is expected that patches with the same neighbourhood topologies would group together, as they share similar visual characteristics.

Our typology reference dataset is formally defined by Kleerekoper (2016). Here, we use a subset of 8 neighbourhood typologies, T (see Table 2). In our selection we consider a variety of different design styles, number of building layers, and construction periods. The typologies are digitized by the climate atlas of the Netherlands initiative (Kleerekoper et al., 2018). This dataset consists of district-level polygons, listing the relative presence (%) of each typology in each district. Since they cover districts, the polygons are only available at a coarser resolution than the grid cells of the LBM. To match the typology presence of the district

Table 2
Neighbourhood typologies considered for our feature vector analyses, as defined by Kleerekoper (2016).

Typology	Period	Characteristics
Historical inner city	<1900	3-5 layers, much concrete
Pre-war block	1900–1940	3-4 layers, moderate amount of greenery
Working-class district	1910–1940	2-3 layers, single-family houses, little to no greenery
Post-war district	1945–1990	2-3 layers, gardens, diversity in housing styles
Cauliflower district	1970–1990	Single-family housing with gardens, winding streets, lots of green
Sub-urban expansion (Vinex)	1990-present	Large diversity in housing styles
Renovated low-rise	1990-present	Neighbourhoods which have undergone renovation
Villas	All	Spacious, single houses

Table 3
Performance difference on the test set between a model trained with only RGB information, and a model with the NIR band included.

Score	RGB-only			RGB+NIR		
	RMSE	R^2	τ	RMSE	R^2	τ
Population	0.045	0.61	0.46	0.051	0.55	0.41
Phys. env	0.049	0.61	0.41	0.05	0.69	0.51
Safety	0.089	0.61	0.50	0.078	0.68	0.47
Amenities	0.043	0.55	0.37	0.041	0.62	0.42
Buildings	0.064	0.70	0.51	0.058	0.73	0.54
Liveability	0.155	0.70	0.52	0.145	0.74	0.54

Table 4
RMSE scores achieved by the model within each built-up area of the test set. (Pop. = population, P.env = physical environment, Amen. = Amenities).

Region	Pop.	P.env	Safety	Amen.	Buildings	Liveability
Dordrecht	0.052	0.048	0.082	0.037	0.067	0.150
Eindhoven	0.044	0.051	0.098	0.038	0.063	0.166
Beesel	0.031	0.046	0.072	0.080	0.047	0.100
Hengelo	0.042	0.048	0.077	0.050	0.063	0.141

Table 5
R^2 scores achieved by the model for each built-up area of the test set. (Pop. = population, P.env = physical environment, Amen. = amenities)

Region	Pop.	P.env	Safety	Amen.	Buildings	Liveability
Dordrecht	0.65	0.47	0.65	0.71	0.76	0.70
Eindhoven	0.66	0.62	0.66	0.57	0.76	0.75
Beesel	0.24	0.31	0.54	0.03	0.60	0.45
Hengelo	0.42	0.56	0.62	0.47	0.65	0.63

level to the grid level, we use the proportion of overlap between the grid cell and each district polygon. For a given typology $t \in T$, a grid cell $g \in G$, and a set of polygons overlapping the grid cell defined as P, we calculate the proportion of each topology present as follows:

$$g_t = \sum_{p=1}^{P} \left(\frac{p_{area}}{g_{area}}\right) p_t \tag{4}$$

3. Results

3.1. Liveability prediction

In Table 3 we show the R^2 and Kendall's τ metrics of both the RGB-only model and the RGB+NIR model on the test set. We show both the five domain scores and the final liveability score, which is regressed directly from the domain scores. The RGB+NIR model is shown to outperform the model with just the RGB bands on most scores, with the notable exception of the population score where a decrease in accuracy occurs. The results show that the addition of NIR information is useful when it is available, as it may result in a better performing model. However, historical aerial images in The Netherlands do not have NIR information. The rest of the results and discussion sections are therefore based on the RGB-only model to maintain compatibility of our analyses with future work.

Table 6
Kendall's τ scores achieved by the model within each built-up area of the test set. (Pop. = population, P.env = physical environment, Amen. = Amenities).

Region	Pop.	P.env	Safety	Amen.	Buildings	Liveability
Dordrecht	0.45	0.40	0.41	0.40	0.56	0.51
Eindhoven	0.49	0.50	0.48	0.38	0.55	0.57
Beesel	0.28	0.32	0.37	−0.02	0.39	0.39
Hengelo	0.39	0.46	0.43	0.36	0.48	0.47

Table 7
Metrics achieved by the model on the validation set and their relative difference to metrics computed over the entire test set.

Score	R^2	% Change	Kendall's τ	% Change
Population	0.84	−26.5%	0.66	−27.4%
Phys. env	0.87	−29.8%	0.64	−21.3%
Safety	0.84	−26.6%	0.65	−36.8%
Amenities	0.95	−41.9%	0.71	−48.9%
Buildings	0.85	−16.4%	0.68	−24.4%
Liveability	0.86	−18.3%	0.67	−22.2%

Table 8
Comparison of our model's overall metrics for the liveability score to an unmodified model tasked with directly predicting liveability from aerial images. The bottleneck model matches an unmodified model in terms of R^2, and surpasses it in Kendall's τ.

Configuration	Val R^2	Test R^2	Val τ	Test τ
Bottleneck	0.861	0.670	0.670	0.521
Baseline	0.801	0.674	0.606	0.484

In Tables 4, 5, and 6 respectively we show the RMSE, R^2 and Kendall's τ metrics obtained by the RGB-only model for each built-up area in our test dataset. Across all regions, our model is able to infer the general trend of all scores, with some noticeable exceptions. Firstly, the achieved metrics can vary strongly per region and domain score, For instance, the model generalizes far less well to Beesel, which is far smaller than the other test sites. However, the decrease in performance is dependent on the domain scores, with some scores being more affected than others.

In Table 7 we show the metrics for the validation set. We also show the difference with the test set to show the capacity of each domain score to generalize to unseen regions. Based on the decrease of metrics between the validation and the test set, our results suggest that *direct* domain scores (*physical environment* and *buildings*) are easier to generalize than *indirect* domain scores. This is mostly the case for buildings (minor decrease in R^2 and even an increase in τ), and to a lesser extent for *physical environment*.

We show a direct comparison between our model with a semantic bottleneck compared to a model which is directly trained to predict liveability in Table 8. Our results show that the use of a bottleneck model mostly improves the performance on this task. While an unconstrained model has a marginally better R^2 score, the bottleneck model outperforms an unconstrained model when considering Kendall's τ.

Lastly, we show the spatial prediction patterns for both the overall liveability score for each test set region, as well as the *buildings* domain score. In Fig. 4 we show the predictions for the *buildings* domain score for all regions in our test set compared to the LBM labels. The patterns

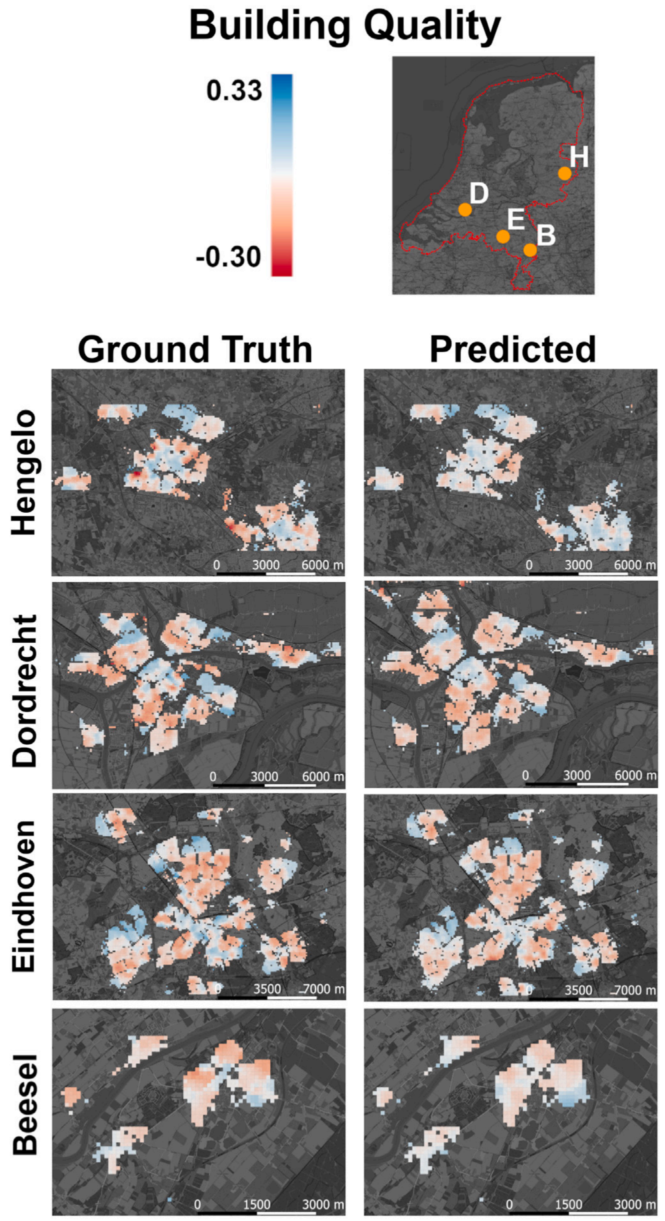

Fig. 4. Predictions for the *buildings* domain score for all regions in the test set. Deeper shades of red represent a low building quality score, while deeper shades of blue denote high building quality. The letters on the left hand side are the first letters of each of our test regions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

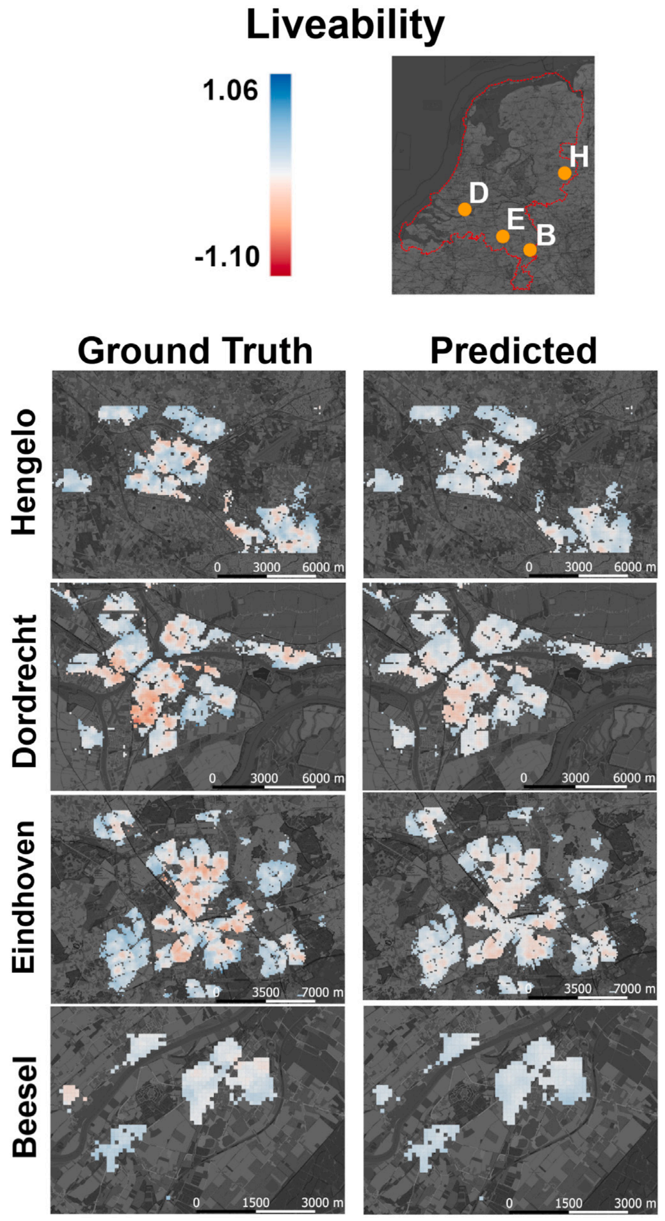

Fig. 5. Predictions for the final liveability score for all regions in the test set. Deeper shades of red represent a lower liveability score, while deeper shades of blue denote a higher patch liveability score. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for the four test region show that our model provides smooth and consistent predictions, and it is able to accurately capture the majority of the fine-grained trends. It is however frequently unable to predict very positive or very negative building quality scores. In Fig. 5 we show the predicted liveability scores for each patch in the test regions. Again, the model predicts the general trend correctly, but struggles to predict values towards either end of the distribution.

3.2. Feature vector analyses

3.2.1. Neighbourhood typologies

In Fig. 6 we show the predicted distribution of scores for each of the neighbourhood types. For each of the selected typologies we show the building quality prediction distributions over the test set. Patches are included when there is 20% or more of the given typology present within the neighbourhood. From these graphs, we show that our model approximates the trend well for most typologies in our unseen test regions, and with a similar accuracy.

In Fig. 7 we show the same plot for the overall liveability score. Trends emerge when comparing the scatterplots for the building quality score to the plots of the overall liveability score. A notable difference is that the model is able to better predict the overall liveability trend of the working-class districts, while it struggles to predict the housing quality of these neighbourhoods in the unseen test regions.

3.2.2. t-SNE embeddings

In Fig. 8 we show a t-SNE plot of the 8 neighbourhood typologies for the global feature vector **r** of Fig. 3, which is the feature vector from which the domain feature matrix C is then derived. The global feature vector **r** therefore represents an aggregate summary of all 5 scores at once. As such, this plot represents which neighbourhood

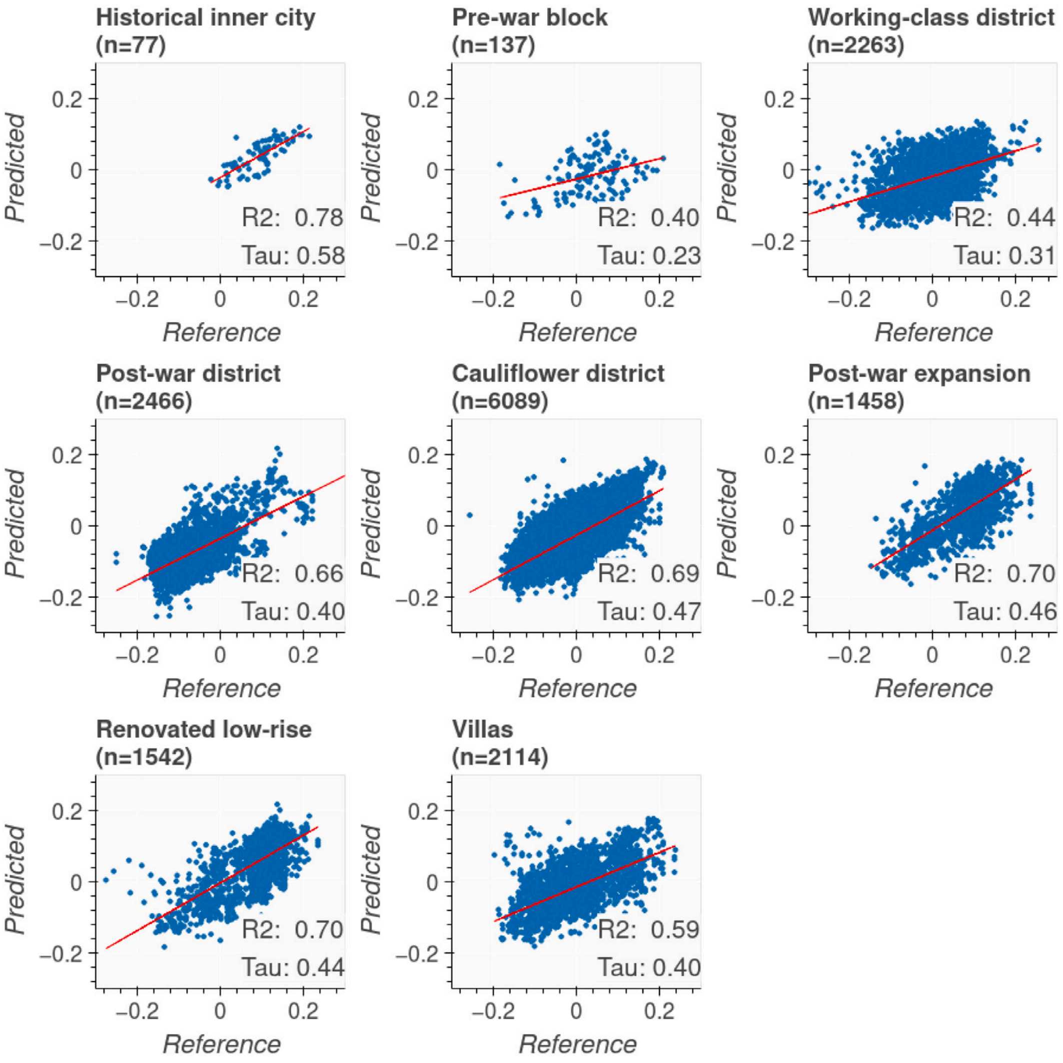

Fig. 6. Scatterplots of the *buildings* domain score for all patches in the test set for each of the neighbourhood typologies considered in this research. Patches are included in a scatterplot when there is 20% or more of the given typology present. We show the reference value of each point on the x-axis, and the predicted value on the y-axis.

typologies are similar across all domain scores. From the graphs we can conclude that most typologies contribute to domain scores in differing ways, resulting a heterogeneous spread across the plots, from which it can be deduced that only a neighbourhood typology as descriptive variable cannot explain the variety of all domain scores. However, it becomes more interesting when we consider the *buildings* domain score using $C_{buildings}$. In Fig. 9 we show a t-SNE plot of the 8 neighbourhood typologies for the *buildings* domain score. This feature vector reflects only how the model perceives the building quality of patches. The plots for the *buildings* domain score reveal that the selected typologies have varying degrees of visual homogeneity, i.e. they occupy different regions of the t-SNE space with different degrees of spread. *Sub-urban expansion* neighbourhoods, *renovated* neighbourhoods, and *historical inner city* neighbourhoods are considered the most visually homogeneous as perceived by our model.

In particular, the *sub-urban expansion* and *renovated district* neighbourhoods form a single cluster of modern building styles (near example 3 of Fig. 10), as both of these typologies only appear after the 1990s. This period saw a paradigm shift towards sub-urban construction, though this cluster does not fully encapsulate sub-urban trends, as for instance villas are still predominantly present outside of it. The top-most cluster in the t-SNE diagram (near example 1 of Fig. 10) shows the dense inner city patterns that are present predominantly in Amsterdam and Rotterdam, both historically and pre-war districts. The visual dissimilarity of these areas from any other building style is particularly striking, as it forms a small but visually distinct cluster while much of the feature space tends to clump together. It shows that these areas have exceptional properties when it comes to building quality. And indeed, when compared to the other cities in the dataset, Amsterdam and Rotterdam are the two most metropolitan areas within the dataset with certain unique features, such as the canal houses in Amsterdam.

4. Discussion

4.1. Predicting the liveability of dutch cities with aerial images and semantic intermediate concepts

The capability of the model to predict various domains varies strongly, as evidenced by Table 7. Between the metrics that have been evaluated, the model is best able to generalize the *direct* domain scores. The *buildings* domain score especially retains good performance for both metrics on the unseen regions. It is followed by the *physical environment* domain score, which sees a greater reduction in the R^2 metric, but retains a high Kendall's τ score. Of the *indirect* domain scores, only the *population* domain score generalizes well to unseen regions. While the *safety* domain score only sees a more drastic reduction in Kendall's τ, the *amenities* domain score sees a dramatic reduction in both R^2 and

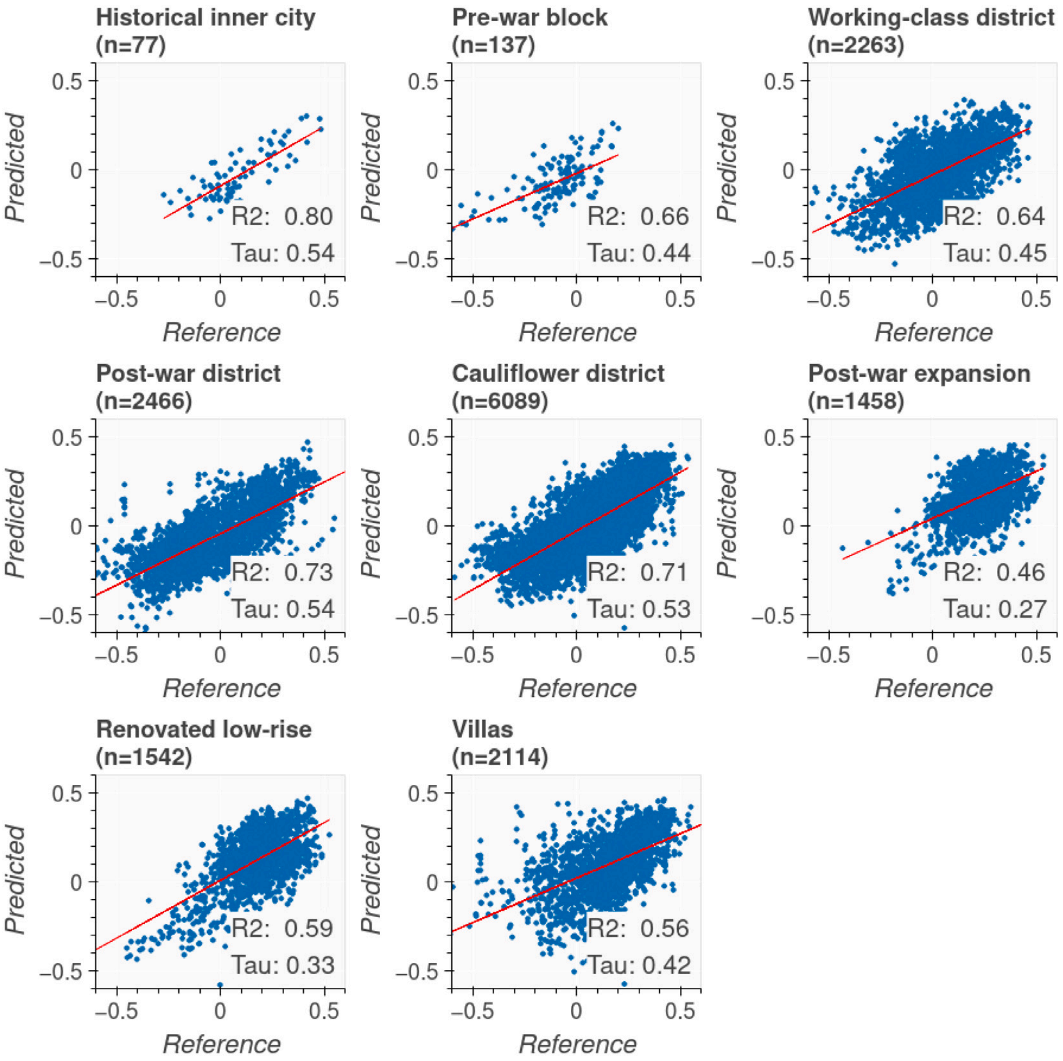

Fig. 7. Scatterplots of the overall liveability score for all patches in the test set for each of the neighbourhood typologies considered in this research. Patches are included in a scatterplot when there is 20% or more of the given typology present. We show the reference value of each point on the x-axis, and the predicted value on the y-axis.

Kendall's τ on the test set. It has the best performance on the validation set, but the strong decrease in performance suggests that amenities are not suitable to predict from aerial images. It should be noted that there are better methods to determine access to amenities compared to prediction from overhead imagery, such as using openly available geodata registries (Sapena et al., 2021). However, in this research we wanted to study the consequences of predicting proxy variables without the use of auxiliary information, in order to study the ensemble of domain scores and their link to liveability in a comprehensive way. Compared to previous literature, our results lead to several observations. Firstly, we corroborate the findings of Arribas-Bel et al. (2017) and Suel et al. (2019) that high-resolution imagery can be used to predict indirect domain information. Secondly, building on Scepanovic et al. (2021), our results also further prove that directly visible domains may be predicted from remote sensing images. Thirdly, we demonstrate that an end-to-end learned regression pipeline from components to liveability (e.g. the two-step regression experiment of Scepanovic et al. (2021) for urban vitality) does not have to come at the cost of performance on the final task. Lastly, our experiments for the first time raise the proposition that domains relating to liveability which are directly predictable from aerial images are easier to generalize to unseen regions than indirect domain scores.

Our results show that the use of an end-to-end trained bottleneck model generally improves model performance to the final task of predicting liveability. Our bottleneck model matches the R^2 metric of the unconstrained baseline model, and slightly surpasses it on Kendall's τ. This shows that a linear mapping from the domain scores (which are a decomposition of the overall liveability score) is sufficient for reconstructing the overall liveability score. The reported metrics corroborate earlier findings that the intermediate prediction of a semantic layer can increase the model's performance of the final task (Levering et al., 2020).

As evidenced by the results, models trained on aerial imagery can transfer fairly well to unseen regions, even across developmental context. The cities in our training dataset have a longer history than two of the cities in our test set, namely Hengelo and Eindhoven. Both of these cities started growing as a result of industrialization. As such, their urban form is partially different than the cities with a longer history. Despite this contrast, our model does not have a decrease in performance compared to Dordrecht, which is a city close to the Rotterdam metropolitan area with a longer history of growth. These results suggest that the learned features are robust between developmental contexts. As a result, our findings suggest that extrapolation of liveability factors to unseen regions is a plausible objective, even when generalizing across developmental contexts. For amenities, the proxy correlations from overhead images is especially tenuous. In the LBM project, the score is originally predicted from distance variables which exceed the size of our 500 m resolution patches, for instance the

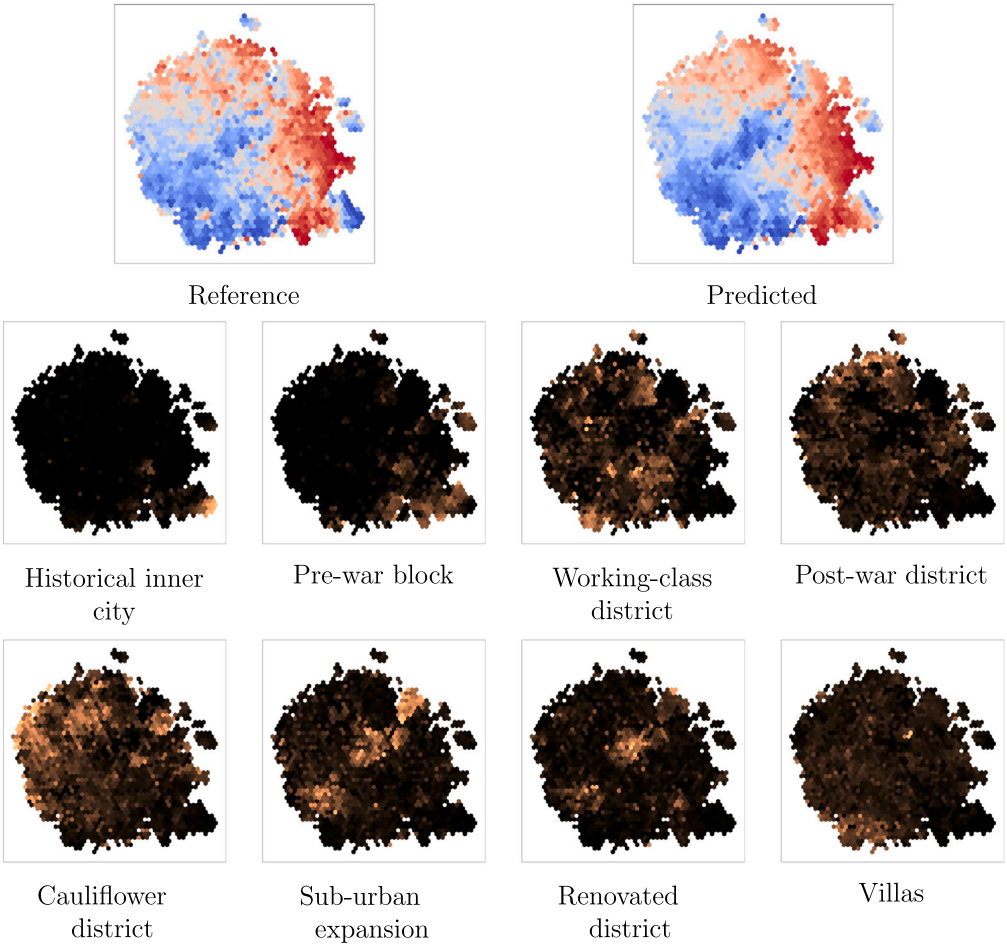

Fig. 8. t-SNE representation of the features from which the domain-specific feature vectors are derived (vector r in Fig. 3), overlaid with the percentage of each typology that is present within a patch. Brighter colours represent a higher percentage of the typology present. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

number of bars within a 2 kilometers radius of the neighbourhood. As such, in a city environment the model can accurately guess that most amenities are close to a neighbourhood. The inclusion of amenities as a dimension score therefore allows us to study proxy variables with an extreme example. The amenities predictions for Beesel in Fig. 11 showcases how transferability becomes a problem with proxy variables that rely on urban context, as the model predictions are not at all correlated with the reference labels. For the other three testing sites this domain score generalizes better, as they are medium to large cities. However, as Beesel is a small village, the model loses geographical context, as the proximity to important amenities such as hospitals is far less certain.

In Section 2.1 we presented the variables used for the domain scores of the LBM project. For the *Population* domain score, the variables included in this domain score can be particularly stigmatizing. The use of ethnicity data has especially drawn criticism from researchers, as the inclusion of ethnicity without accounting for confounding variables may lead to false stereotypes. While not accounting for confounding variables, the first version LBM was already used to justify policy decisions. The main concern for the research was to maximize the R^2 coefficient, and as a result the researchers did not take into account the importance of mitigating stigmatism (Uitermark et al., 2017). The second version of the LBM has attempted to mitigate the stigmatizing effects of including ethnicity variables by only using the residuals after accounting for income. However, it was still widely criticised (Baggerman, 2020; Teeffelen, 2021). In version 3.0 of the LBM, stigmatizing variables such as the ones used by population score have been phased out in favour of a more generalized domain, namely *social cohesion* (Leidelmeijer and Mandemakers, 2020). However, during our analysis this improved version was not yet available. In our research we have decided to use the population score as it represents a generalized score which allows us to determine how well socio-economic and socio-demographic data can be predicted, and how well this domain will generalize to unseen regions. However, we refrain from analysing prediction patterns for this domain score so as to not perpetuate or justify the use of these stigmatizing variables.

4.2. Perspectives for liveability monitoring with EO

In this section we discuss how our research can help to provide deeper perspectives for liveability monitoring from remotely sensed imagery. There are several possible approaches for modelling liveability using remote sensing. When liveability reference data is not available, deterministic intermediate variables may be used as a proxy, where it is assumed that each variable is an indicator for liveability. For instance, the United Kingdom uses an *index of multiple deprivation*, which measures the relative deprivation of *Lower Layer Super Output* areas with a mean population of 1,500 residents. The index is measured at a fine-grained neighbourhood scale, and considers 7 domains (Income, Employment, Education, Health, Crime, Barriers to Housing and Services, and Living Environment) (Penney, 2019). It combines these dimensions into a final deprivation index using different weights for each dimension using guidance from liveability theory. Some of the individual deprivation factors of this index of multiple deprivations

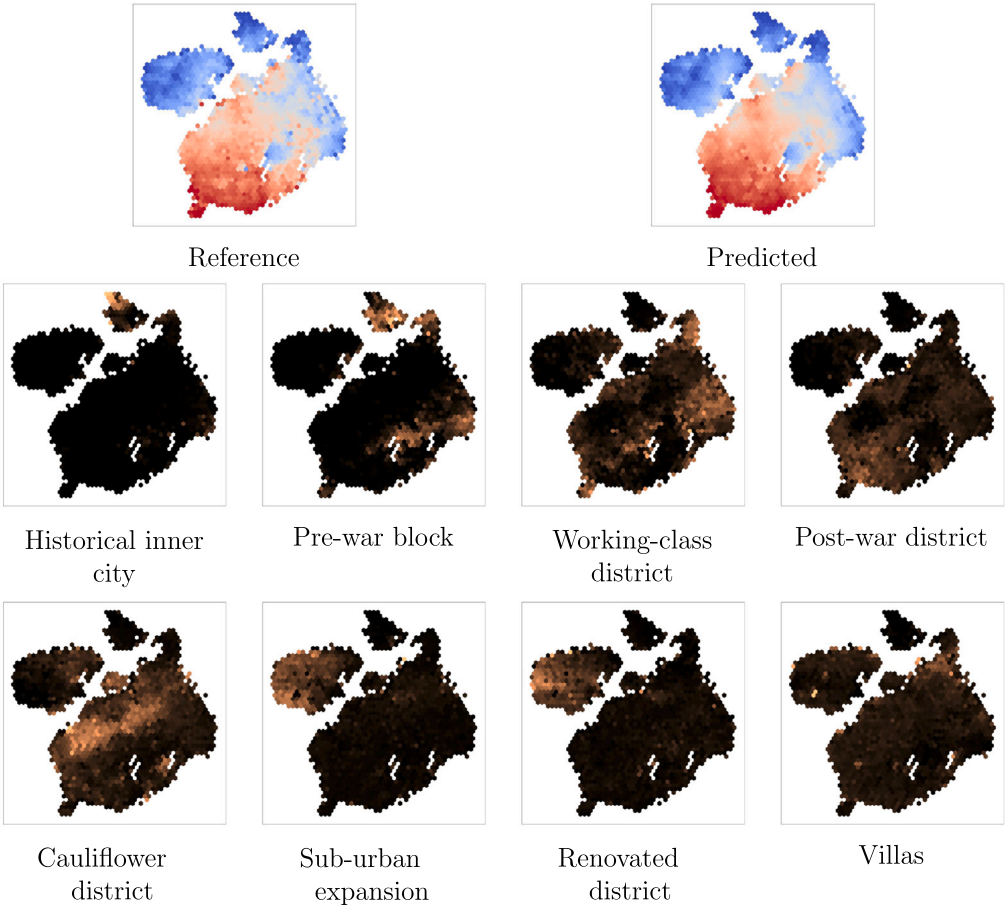

Fig. 9. t-SNE representation of the features used to predict the *buildings* domain score, overlaid with the percentage of each typology that is present within a patch. Brighter colours represent a higher percentage of the typology present. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

have been predicted through remote sensing (Arribas-Bel et al., 2017; Suel et al., 2021). In such a set-up, the role of remote sensing would be to interpolate and extrapolate intermediate variables in support of liveability modelling. Liveability can also be modelled through remote sensing in an end-to-end manner by first predicting intermediate factors, and then to recombine them into a liveability score by using expert opinions (Huang and Liu, 2022). Such a method allows for the acquisition of large-scale inventories of liveability measurements without needing any reference data. The downside to this deterministic measuring process is that the importance of intermediate variables to liveability is not calibrated empirically through resident opinions. As such, the expert opinions on which intermediate variables matter most may be different from the liveability as experienced by residents.

As a compromise between deterministic and empirical liveability modelling, hedonic pricing assumes that housing prices are in part indicative of the liveability of a neighbourhood, as people are willing to pay more for houses in liveable areas. This is a simple and scalable assumption, which makes it attractive for large-scale modelling, as the definition of the reference data is constant no matter the location. Therefore, if house sales information is available it may serve as a proxy for the liveability of an area (Bency et al., 2017; Yao et al., 2018). However, the downside of hedonic pricing is that the assumed contribution of liveability is not tested against resident opinions either, meaning that it may still be off from the liveability experienced by residents. Moreover, a model may need more information to infer how much signal can be attributed to the desireability of a location. For instance, an area may have a poor quality of the built environment, but very attractive surroundings. As a result it may trend to the average.

The most informative type of reference data is based on surveyed resident opinions. Such a data source does not assume that there is a relation between proxy factors and the experienced liveability, but provides the evidence to directly test such hypothesis in practice. However, fine-grained liveability reference data based on residents' opinions are only scarcely available. Surveying efforts are expensive, hard to perform on a large scale, and sensitive to a variety of biases such as response bias (the tendency for respondents to give inaccurate answers) and participation bias (the inability or unwillingness of certain groups of residents to respond). A well-performed study at scale is therefore a labour-intensive process. The privacy of respondents also needs to be respected, which further complicates the spatial scale at which information is typically reported. This makes the LBM a remarkable project, as it is on a fine spatial scale and is partially modelled on the subjective opinions of residents. To our knowledge, it provides the first large-scale yet fine-grained dataset of liveability which incorporates resident opinions, thus opening possibilities for understanding liveability in the Dutch context, but with limitations that we discuss in the next Section.

4.3. Limitations

In our research we use $1m$ resolution aerial image patches from the nationally-available aerial image, which is open data, while the liveability reference data is of a fine spatial resolution and nation-wide available as well. The unique availability of both data sources allows us to observe liveability with an unprecedented geographical scale, resolution, and fidelity. While this work does allow to pursue the limits of what may be measured at scale from remote sensing imagery, it restrains the methodology to regions with a similar data availability.

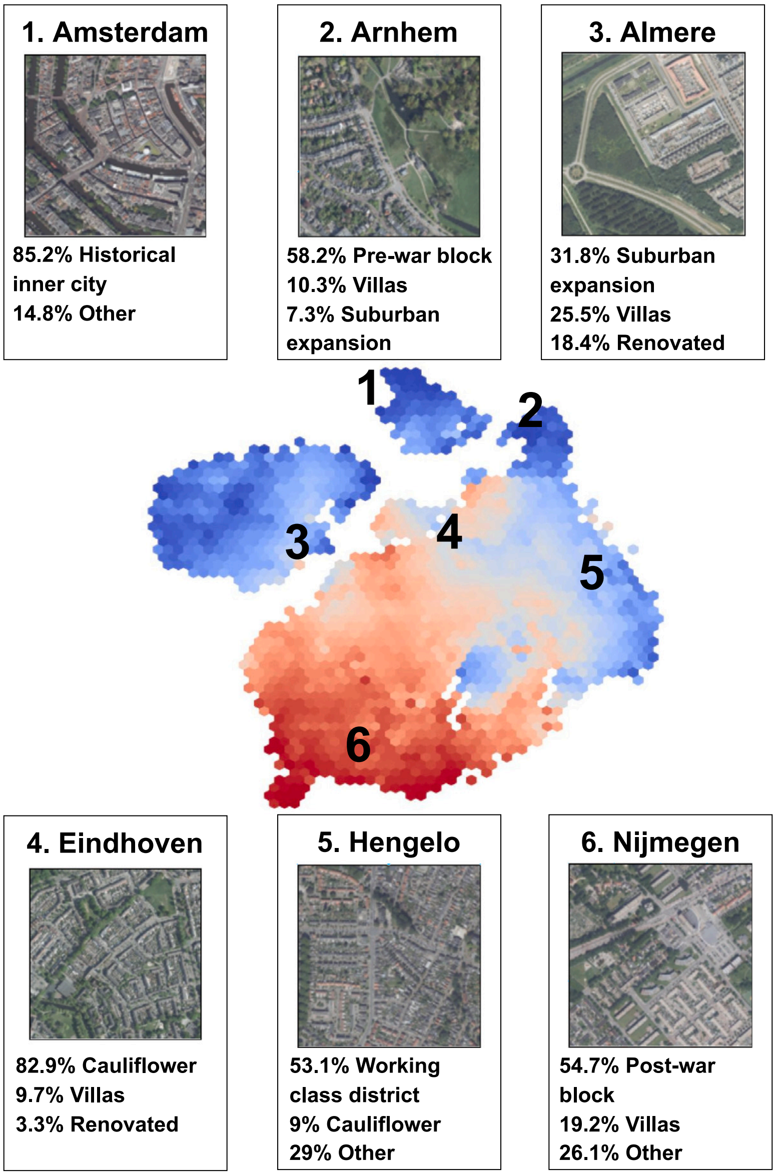

Fig. 10. Example aerial image patches with their corresponding neighbourhood-level typology labels plotted over the *buildings* domain score embedding. Note that the neighbourhood typology information is often only available at a coarser spatial scale and therefore they may not fully represent the individual patch content.

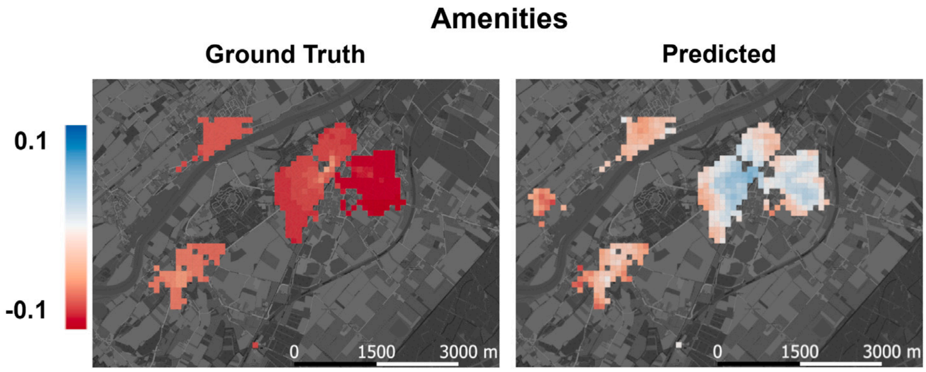

Fig. 11. Predictions for the Amenities domain score over the region of Beesel. The maps highlight that the model fails to predict the trends present in the villages, leading to negative performance metrics for this test region.

Despite this restriction, our results may be replicated in other countries with liveability labels through commercial satellite services. In that sense, our results are scalable to any region, but most strongly applicable to regions with a high data availability. Our results indicate that domains that can be observed directly through aerial imagery are easier to generalize than domains which need proxies in order to be predicted. This has practical implications for using remote sensing to fill gaps in data availability for the purposes of predicting liveability. For instance, where possible amenity data should be derived from sources other than remote sensing imagery, as open geodata registries provide coverage of the most important amenities for most countries. However, if building types are largely homogeneous between two areas but building quality data is only available in one area, then it may be worthwhile to gap-fill this data through remote sensing.

While liveability monitoring from Earth observation has been proven to work for several different research cases and for each of the different types of reference datasets that are available, the subjectivity of the topic continues to hamper comparisons across studies. First, there is no standard definition for liveability, which plays a role in determining a common ground for liveability studies (Paul and Sen, 2020). Second, the way it is measured varies for each study, as do the variables and methods used to measure liveability. We therefore consider liveability prediction to only be valid within the cultural context in which it is measured, with very limited generalization beyond this context. In other words, the values that make a place liveable are culture- and location-specific. As such, we do not believe that liveability prediction models could be applied out of the box in a completely different cultural context. While our models retain sufficient performance in unseen cities for the extrapolation of liveability surveys, the entirety of our dataset falls within the same cultural context, which is the Netherlands as a country. As such, our dataset has a largely homogeneous cultural and policy context. Attempting to extrapolate outside of the Netherlands, e.g. attempting to predict liveability in Belgium or Germany with our model, will most likely be less successful, due to a difference in cultural and policy context.

The LBM project is an ongoing project which is still being updated. While the input variables and the domains are updated between versions, the reference data upon which the liveability scores are calculated remains unchanged. Meanwhile, the aerial image data will be updated yearly for the foreseeable future. As such, the data used in our study can theoretically be used to test whether the relation between the spatial configuration of settlements and their liveability is persistent over time. As the temporal extent of the datasets increase throughout the years, this option will become more salient as significant changes to the liveability of a neighbourhood such as gentrification and impoverishment will take years to manifest.

5. Conclusions

In this paper we study the prediction of liveability from aerial images at the neighbourhood level for 13 built-up areas in the Netherlands. To do so, we test the applicability of remote sensing to predict five domain scores relating to liveability. We assess how well domains that can be learned directly from the image content itself (*physical environment* and *buildings*) can be predicted, as well as domains which require proxy correlations (*population*, *safety*, and *amenities*). Our results indicate that liveability domain scores generalize fairly well to unseen regions, even in regions which have a different developmental context. Furthermore, our results indicate that domains which can be directly predicted from the image pixels generalize better than domains which rely on proxy correlations, as the reduction in performance between the validation and the test set is lower for these predicted domain scores. We also study how our model perceives the liveability of different neighbourhood typologies. Our results indicate that our model is proficient at recognizing the liveability of different urban typologies, though with varying accuracy. Secondly, through t-SNE dimensionality reduction we inferred how our model observes homogeneity within neighbourhood typologies. Our results show that our model considers certain neighbourhood typologies to be visually distinct for the purposes of recognizing building quality, but less so for overall liveability. Our research suggests that remote sensing can be used to extrapolate liveability surveys to new and unseen regions within the same cultural and policy context. Finally, our study may enable longitudinal studies across time series of aerial images in order to monitor liveability. The code for our project is available at https://github.com/Bixbeat/liveability-rs

CRediT authorship contribution statement

Alex Levering: Conceptualization, Software, Methodology, Visualization, Writing – original draft. **Diego Marcos:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Jasper van Vliet:** Conceptualization, Writing – review & editing. **Devis Tuia:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset and the code for reproducing our model are available at https://github.com/ahlevering/liveability-rs

References

Arribas-Bel, D., Patino, J.E., Duque, J.C., 2017. Remote sensing-based measurement of Living Environment Deprivation: Improving classical approaches with machine learning. PLOS ONE (ISSN: 1932-6203) 12 (5), e0176684. http://dx.doi.org/10.1371/journal.pone.0176684, URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176684. Publisher: Public Library of Science.

Baggerman, K., 2020. Migratieachtergrond? Volgens de Leefbaarometer maak jij je wijk dan slechter. URL: https://stadszaken.nl/artikel/2804/migratieachtergrond-volgens-de-leefbaarometer-maak-jij-je-wijk-dan-slechter.

Barber, S., Hickson, D.A., Kawachi, I., Subramanian, S.V., Earls, F., 2016. Neighborhood Disadvantage and Cumulative Biological Risk Among a Socioeconomically Diverse Sample of African American Adults: An Examination in the Jackson Heart Study. J. Racial Ethn. Health Dispar. (ISSN: 2196-8837) 3 (3), 444–456. http://dx.doi.org/10.1007/s40615-015-0157-0.

Bechtel, B., Alexander, P., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L., Stewart, I., 2015. Mapping Local Climate Zones for a Worldwide Database of the Form and Function of Cities. ISPRS Int. J. Geo-Inf. 4, 199–219. http://dx.doi.org/10.3390/ijgi4010199.

Bency, A.J., Rallapalli, S., Ganti, R.K., Srivatsa, M., Manjunath, B.S., 2017. Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 320–329. http://dx.doi.org/10.1109/WACV.2017.42.

CBS, 2016. Inkomen per gemeente en wijk, 2016. URL: https://www.cbs.nl/nl-nl/maatwerk/2019/02/inkomen-per-gemeente-en-wijk-2016. Last Modified: 2019-01-18T00:00:00+01:00.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. (ISSN: 1063-6919) pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.

Evans, G.W., 2003. The built environment and mental health. J. Urban Health: Bull. N Y Acad. Med. (ISSN: 1099-3460) 80 (4), 536–555. http://dx.doi.org/10.1093/jurban/jtg063.

Haan, M., Kaplan, G.A., Camacho, T., 1987. Poverty and health. Prospective evidence from the Alameda County Study. Am. J. Epidemiol. (ISSN: 0002-9262) 125 (6), 989–998. http://dx.doi.org/10.1093/oxfordjournals.aje.a114637.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: CVPR. pp. 770–778.

Huang, X., Liu, Y., 2022. Livability assessment of 101,630 communities in China's major cities: A remote sensing perspective. Sci. China Earth Sci. (ISSN: 1869-1897) 65 (6), 1073–1087. http://dx.doi.org/10.1007/s11430-021-9896-4.

Jensen, R., Gatrell, J., Boulton, J., Harper, B., 2004. Using Remote Sensing and Geographic Information Systems to Study Urban Quality of Life and Urban Forest Amenities. Ecol. Soc. (ISSN: 1708-3087) 9 (5), URL: https://www.jstor.org/stable/26267693. Publisher: Resilience Alliance Inc..

van Kamp, I., Leidelmeijer, K., Marsman, G., de Hollander, A., 2003. Urban environmental quality and human well-being: Towards a conceptual framework and demarcation of concepts; a literature study. Landsc. Urban Plan. (ISSN: 0169-2046) 65 (1), 5–18. http://dx.doi.org/10.1016/S0169-2046(02)00232-3, URL: https://www.sciencedirect.com/science/article/pii/S0169204602002323.

Kendall, M.G., 1938. A New Measure for Rank Correlation. Biometrika (ISSN: 0006-3444) 30 (1–2), 81–93. http://dx.doi.org/10.1093/biomet/30.1-2.81, URL: https://doi.org/10.1093/biomet/30.1-2.81.

Kleerekoper, L., 2016. Urban Climate Design: Improving thermal comfort in Dutch neighbourhoods. A+BE: Archit. Built Environ. 6, http://dx.doi.org/10.7480/abe.2016.11.

Kleerekoper, L., Koekoek, A., Kluck, J., 2018. Een wijktypologie voor klimaatadaptatie. Stad. Mag. 28–30, URL: https://www.hva.nl/binaries/content/assets/subsites/kc-techniek/publicaties-klimaatbestendige-stad/kleerekoper_2018_sw01_wijktypologie.pdf?1518015946804.

Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P., 2020. Concept Bottleneck Models. In: Proceedings of the 37th International Conference on Machine Learning. PMLR, pp. 5338–5348, URL: https://proceedings.mlr.press/v119/koh20a.html. iSSN: 2640-3498.

Kuffer, M., Thomson, D.R., Boo, G., Mahabir, R., Grippa, T.s., Vanhuysse, S., Engstrom, R., Ndugwa, R., Makau, J., Darin, E., de Albuquerque, J.P., Kabaria, C., 2020. The Role of Earth Observation in an Integrated Deprived Area Mapping "System" for Low-to-Middle Income Countries. Remote Sens. (ISSN: 2072-4292) 12 (6), 982. http://dx.doi.org/10.3390/rs12060982, URL: https://www.mdpi.com/2072-4292/12/6/982. number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

Leidelmeijer, K., Mandemakers, J., 2020. Leefbaarheid in Nederland 2020. Technical Report, Atlas Research, Amsterdam, p. 79.

Leidelmeijer, K., Marlet, G., Ponds, R., Schulenberg, R., van Woerkens, C., 2014. Leefbaarometer 2.0: Instrumentenontwikkeling. Technical Report 2, RIGO Research en Advies/Atlas voor Gemeenten, Amsterdam, p. 151, URL: https://doc.leefbaarometer.nl/resources/Leefbaarometer%202.0%20Instrumentontwikkeling.pdf.

Levering, A., Marcos, D., Lobry, S., Tuia, D., 2020. Interpretable Scenicness from Sentinel-2 Imagery. In: Proceedings of the 2020 International Geoscience and Remote Sensing Symposium. Hawaii, p. 4.

Li, G., Weng, Q., 2007. Measuring the quality of life in city of Indianapolis by integration of remote sensing and census data. Int. J. Remote Sens. (ISSN: 0143-1161) 28 (2), 249–267. http://dx.doi.org/10.1080/01431160600735624, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01431160600735624.

Liu, S., Shi, Q., 2020. Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China. ISPRS J. Photogramm. Remote Sens. (ISSN: 0924-2716) 164, 229–242. http://dx.doi.org/10.1016/j.isprsjprs.2020.04.008, URL: https://www.sciencedirect.com/science/article/pii/S0924271620301052.

Loshchilov, I., Hutter, F., 2019. Decoupled Weight Decay Regularization. In: ICLR.

Maaten, L.v.d., Hinton, G., 2008. Visualizing Data using t-SNE. J. Mach. Learn. Res. (ISSN: 1533-7928) 9 (86), 2579–2605, URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Marcos, D., Fong, R., Lobry, S., Flamary, R., Courty, N., Tuia, D., 2021. Contextual Semantic Interpretability. In: Ishikawa, H., Liu, C.-L., Pajdla, T., Shi, J. (Eds.), Computer Vision – ACCV 2020. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, ISBN: 978-3-030-69538-5, pp. 351–368. http://dx.doi.org/10.1007/978-3-030-69538-5_22.

Ministry of Infrastructure and the Environment, 2012. 35 icons of Dutch spatial planning. Technical Report, Ministry of Infrastructure and the Environment, The Hague, URL: https://open.overheid.nl/repository/ronl-archief-a0559ae0-6613-411f-bd8b-aaf7f1d80a6f/1/pdf/ro-35-icons.pdf.

Paul, A., Sen, J., 2020. A critical review of liveability approaches and their dimensions. Geoforum; J. Phys. Hum. Reg. Geosci. (ISSN: 0016-7185) 117, 90–92. http://dx.doi.org/10.1016/j.geoforum.2020.09.008, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7529590/.

PDOK, 2017. NIEUW: hogere resolutie luchtfoto als open data bij PDOK. URL: https://www.pdok.nl/-/nieuw-hogere-resolutie-luchtfoto-als-open-data-bij-pdok.

Penney, B., 2019. English indices of deprivation 2015. Technical Report, UK Office for National Statistics.

Qiu, C., Mou, L., Schmitt, M., Zhu, X.X., 2019. Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network. ISPRS J. Photogramm. Remote Sens. (ISSN: 0924-2716) 154, 151–162. http://dx.doi.org/10.1016/j.isprsjprs.2019.05.004, URL: https://www.sciencedirect.com/science/article/pii/S0924271619301261.

Rahman, A., Kumar, Y., Fazal, S., Bhaskaran, S., 2011. Urbanization and Quality of Urban Environment Using Remote Sensing and GIS Techniques in East Delhi-India. J. Geogr. Inf. Syst. 03 (01), 62. http://dx.doi.org/10.4236/jgis.2011.31005, URL: http://www.scirp.org/journal/PaperInformation.aspx?PaperID=3707&#abstract. number: 01 Publisher: Scientific Research Publishing.

Rodriguez Lopez, J.M., Heider, K., Scheffran, J., 2017. Frontiers of urbanization: Identifying and explaining urbanization hot spots in the south of Mexico City using human and remote sensing. Appl. Geogr. (ISSN: 0143-6228) 79, 1–10. http://dx.doi.org/10.1016/j.apgeog.2016.12.001, URL: https://www.sciencedirect.com/science/article/pii/S0143622816307421.

Rosier, J.F., Taubenböck, H., Verburg, P.H., van Vliet, J., 2022. Fusing Earth observation and socioeconomic data to increase the transferability of large-scale urban land use classification. Remote Sens. Environ. (ISSN: 0034-4257) 278, 113076. http://dx.doi.org/10.1016/j.rse.2022.113076, URL: https://www.sciencedirect.com/science/article/pii/S0034425722001900.

Sapena, M., Wurm, M., Taubenböck, H., Tuia, D., Ruiz, L., 2021. Estimating quality of life dimensions from urban spatial pattern metrics. Comput. Environ. Urban Syst. 85, 101549. http://dx.doi.org/10.1016/j.compenvurbsys.2020.101549.

Scepanovic, S., Joglekar, S., Law, S., Quercia, D., 2021. Jane Jacobs in the Sky: Predicting Urban Vitality with Open Satellite Data. ACM Hum.-Comput. Interact. 5, 1–25. http://dx.doi.org/10.1145/3449257.

Singleton, A., Arribas-Bel, D., Murray, J., Fleischmann, M., 2022. Estimating generalized measures of local neighbourhood context from multispectral satellite images using a convolutional neural network. Comput. Environ. Urban Syst. (ISSN: 0198-9715) 95, 101802. http://dx.doi.org/10.1016/j.compenvurbsys.2022.101802, URL: https://www.sciencedirect.com/science/article/pii/S0198971522000461.

Srivastava, S., Vargas-Muñoz, J.E., Tuia, D., 2019. Understanding urban landuse from the above and ground perspectives: a deep learning, multimodal solution. Remote Sensing of Environment (ISSN: 00344257) 228, 129–143. http://dx.doi.org/10.1016/j.rse.2019.04.014, https://www.sciencedirect.com/science/article/pii/S0034425719301579.

Suel, E., Bhatt, S., Brauer, M., Flaxman, S., Ezzati, M., 2021. Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. Remote Sens. Environ. (ISSN: 0034-4257) 257, 112339. http://dx.doi.org/10.1016/j.rse.2021.112339, URL: https://www.sciencedirect.com/science/article/pii/S0034425721000572.

Suel, E., Polak, J.W., Bennett, J.E., Ezzati, M., 2019. Measuring social, environmental and health inequalities using deep learning and street imagery. Sci. Rep. (ISSN: 2045-2322) 9 (1), 6229. http://dx.doi.org/10.1038/s41598-019-42036-w, URL: https://www.nature.com/articles/s41598-019-42036-w. number: 1 Publisher: Nature Publishing Group.

Taubenböck, H., Esch, T., Felbier, A., Wiesner, M., Roth, A., Dech, S., 2012. Monitoring urbanization in mega cities from space. Remote Sensing of Urban Environments, Remote Sens. Environ. (ISSN: 0034-4257) Remote Sensing of Urban Environments, 117, 162–176. http://dx.doi.org/10.1016/j.rse.2011.09.015.URL: https://www.sciencedirect.com/science/article/pii/S0034425711003427,

Teeffelen, K.v., 2021. Een algoritme is niet neutraal, ook een overheidsalgoritme niet. Trouw URL: https://www.trouw.nl/gs-bbc021d0.

Thompson, S., Kent, J., 2014. Healthy Built Environments Supporting Everyday Occupations: Current Thinking in Urban Planning. J. Occup. Sci. (ISSN: 1442-7591) 21 (1), 25–41. http://dx.doi.org/10.1080/14427591.2013.867562.

Tian, Y., Tsendbazar, N.-E., van Leeuwen, E., Fensholt, R., Herold, M., 2022. A global analysis of multifaceted urbanization patterns using Earth Observation data from 1975 to 2015. Landsc. Urban Plan. (ISSN: 0169-2046) 219, 104316. http://dx.doi.org/10.1016/j.landurbplan.2021.104316, URL: https://www.sciencedirect.com/science/article/pii/S0169204621002796.

Uitermark, J., Hochstenbach, C., van Gent, W., 2017. The statistical politics of exceptional territories. Polit. Geogr. (ISSN: 0962-6298) 57, 60–70. http://dx.doi.org/10.1016/j.polgeo.2016.11.011, URL: https://www.sciencedirect.com/science/article/pii/S096262981630258X.

Veenhoven, R., Ehrhardt, J., Ho, M.S.D., de Vries, A., 1993. Happiness in nations: Subjective appreciation of life in 56 nations 1946–1992. In: HappinEss in Nations: subjective Appreciation of Life in 56 Nations 1946–1992. Erasmus University Rotterdam, Rotterdam, Netherlands, ISBN: 978-90-72597-46-5, p. 365.

Yao, Y., Zhang, J., Hong, Y., Liang, H., He, J., 2018. Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. Trans. GIS (ISSN: 1467-9671) 22 (2), 561–581. http://dx.doi.org/10.1111/tgis.12330, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12330. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12330.