Contrastive Pretraining for Visual Concept Explanations of Socioeconomic Outcomes

Ivica Obadic, Alex Levering, Lars Pennig, Dario Oliveira, Diego Marcos, Xiaoxiang Zhu

Contrastive Pretraining for Visual Concept Explanations of Socioeconomic Outcomes

Ivica Obadic[1,2] Alex Levering[3] Lars Pennig[1] Dario Oliveira[4] Diego Marcos[5] Xiaoxiang Zhu[1,2]
[1]Technical University of Munich [2]Munich Center for Machine Learning [3]Vrije Universiteit Amsterdam
[4]Getulio Vargas Foundation [5] Inria, University of Montpellier

Abstract

Predicting socioeconomic indicators from satellite imagery with deep learning has become an increasingly popular research direction. Post-hoc concept-based explanations can be an important step towards broader adoption of these models in policy-making as they enable the interpretation of socioeconomic outcomes based on visual concepts that are intuitive to humans. In this paper, we study the interplay between representation learning using an additional task-specific contrastive loss and post-hoc concept explainability for socioeconomic studies. Our results on two different geographical locations and tasks indicate that the task-specific pretraining imposes a continuous ordering of the latent space embeddings according to the socioeconomic outcomes. This improves the model's interpretability as it enables the latent space of the model to associate urban concepts with continuous intervals of socioeconomic outcomes. Further, we illustrate how analyzing the model's conceptual sensitivity for the intervals of socioeconomic outcomes can shed light on new insights for urban studies.

1. Introduction

Monitoring socioeconomic outcomes is crucial for effective policy-making and progress toward the Sustainable Development Goals (SDGs). An emerging approach in recent years is to predict these indicators by training deep learning models on satellite imagery [7]. For example, such approaches are applied for estimating the economic well-being in Africa [23, 31], the income in France [1], the urban vitality in Italy [25], the landscape scenicness in the UK [20], or the livability of cities in the Netherlands [21]. However, these approaches are rarely used in practice by decision-makers [7], with one of the main obstacles being the opaque nature of the deep learning models, whose inner workings and mechanisms lack human interpretability [14].

The interpretability challenges in the proposed deep learning models for socioeconomic indicator estimation have usually been addressed with saliency map methods like Grad-CAM [1] or concept bottleneck models [20, 21, 25]. While saliency maps identify relevant regions in the image by highlighting raw pixels, they do not provide additional insights into the high-level semantics the model extracted for its inference [2]. Furthermore, they are of questionable reliability as they can be insensitive to the changes in the model weights or data distribution [3]. A promising alternative to the saliency maps methods are the concept-based explanations that relate a target variable to a set of explanatory concepts. Compared to the saliency maps, they more closely resemble human reasoning and thus are more intuitive [18]. Moreover, concept-based explanations are a natural fit for socioeconomic studies, as concepts are commonly used as indicator variables [19]. These approaches for remote sensing have only been attempted through bottleneck models, which relate intermediate concepts to an output variable in an end-to-end trained manner [20, 21, 25]. However, such methods also have several downsides. They require that a dataset of annotated concepts specific to the study region is available during training to condition the bottleneck and assume a strong relation between the concepts and the target variables. The post-hoc concept-based explanations can overcome these issues as they relate the outcomes to a set of human interpretable concepts after the model training phase [12, 17, 34]. These methods identify the concepts by looking for a direction in the latent space of the trained model towards which the concept examples are found. The concepts are represented with coherent images sharing similar visual characteristics; thus, the resulting explanations are intuitively understandable [12, 17]. Post-hoc concept explanation methods have already been successfully applied in other sensitive domains like medicine [13, 17]. These approaches can also be beneficial for remote sensing studies, as, unlike concept bottleneck models, they do not require an additional set of concept labels specific to the geographical area of the study region.

In this work, we introduce a new expandability pipeline (visualized in Figure 1) that uses post-hoc concept explanations to reveal how deep learning models predict socioeco-

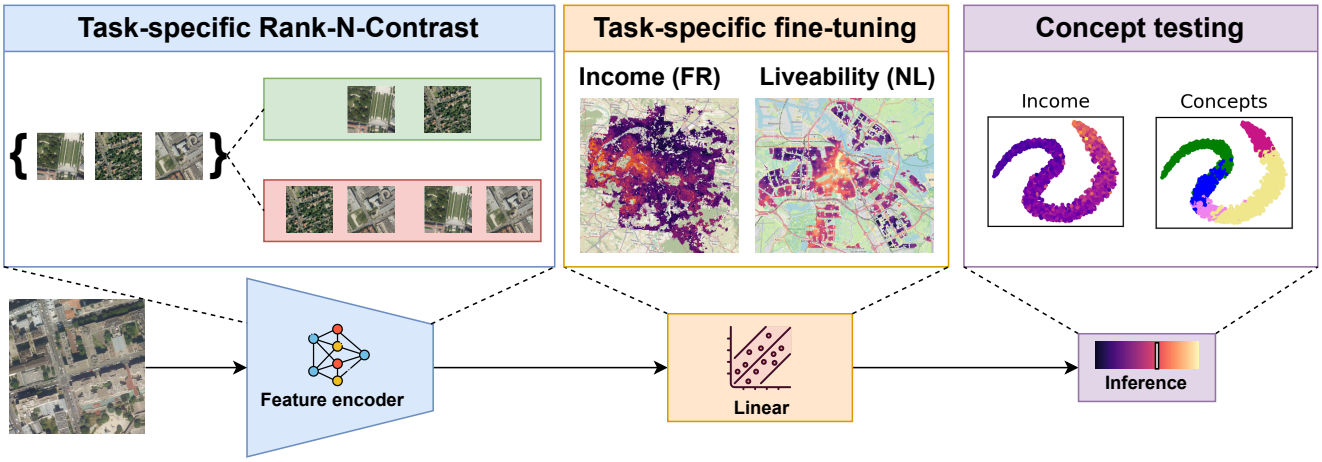

Figure 1. **Flowchart of the pipeline presented in this research.** For both tasks, we use Rank-n Contrast to pre-train the feature encoder to produce embeddings that strongly relate to socioeconomic task scores. Secondly, we freeze the encoder weights and probe a linear layer to regress the task-specific score. Lastly, we use TCAV to study the relation of various urban concepts to socioeconomic scores.

nomic outcomes in terms of easy-to-understand urban concepts. It consists of three main steps: 1) a task-specific contrastive pre-training of a deep learning encoder to continuously order the latent space embeddings based on their socioeconomic outcome, 2) probing a linear layer on the top of the encoder to predict socioeconomic outcome, and 3) concept testing with the Testing with Concept Activation Vectors (TCAV) method [17]. We evaluate our pipeline on two tasks for estimating socioeconomic indicators from remote sensing imagery, namely income prediction in France [1], and liveability estimation in the Netherlands [21]. In summary, our work evaluates for the first time visual post-hoc concept-based explanations in remote sensing socioeconomic studies with the following main contributions:

- Continuously ordering the latent space embeddings based on their socioeconomic outcomes yields comparable or improved prediction performance to a baseline supervised learning.
- Such latent space improves the interpretability of the deep learning model as the concepts cluster according to continuous intervals of socioeconomic outcomes.
- Finally, this enables studying the sensitivities of socioeconomic outcomes to different urban concepts.

2. Related Work

Although essential for their wider adoption in practice, the interpretability of the deep learning models for socioeconomic indicator estimation is rarely tackled [14]. The approaches in related studies mainly propose using saliency maps [1] or concept bottleneck models [20, 21, 25]. Concretely, Abitbol and Karsai trains an EfficientNet deep learning model on aerial imagery to predict household income in five French cities [1]. They used the Grad-CAM saliency map method to infer the relevant regions for the model predictions. The identified regions are matched with the urban class polygons to reveal the relationship between urban classes and the income prediction of the deep learning model. Their results show that residential areas can both activate in case of high or low socioeconomic status depending on the city and area, whereas the presence of infrastructure such as motor- or railways is consistently related to a lower household income. When it comes to concept bottleneck models, they are used by Scepanovic et al. to predict the urban vitality in six Italian cities from Sentinel-2 imagery [25]. In the first step, the authors use a VGG-16 network and a convolutional autoencoder to predict six different land use and building block characteristics from image patches. These features are then used as input to the linear model that estimates urban vitality. The authors discover, that intersection density, small parks, anisotropicity, and land use mix positively influence vitality, while the variables block size and building height negatively impact the vitality. Further, Levering et al. train an end-to-end convolutional neural network with a semantic bottleneck to predict the landscape scenicness in the UK based on Sentinel-2 imagery [20]. The interpretable concepts in the semantic bottleneck layer encode the region's land cover distribution. This architecture enables an explanation of the predicted scenicness score in terms of the land cover characteristics of the region. The authors find, that scenic areas are often associated with mountainous, coastal regions, while urban areas and the presence of human influence are negatively related to the landscape scenicness. The semantic bottleneck models are also used in [21] to predict the livability in

the Dutch cities using aerial imagery. Again, the authors include a semantic bottleneck layer, which predicts five different intermediate domain scores. Their results indicate, that some intermediate domains such as buildings or physical environments are much better suited to be predicted from aerial imagery than others like safety or amenities.

While the above-mentioned approaches shed light on the relevant factors for socioeconomic indicator prediction, they rely on the availability of labeled features specific to the study region. Hence, the limited data availability in remote sensing represents an obstacle to applying these approaches to different geographical areas. Further, saliency maps are known to produce local, low-level feature explanations with limited faithfulness [3]. In this work, we aim to overcome these limitations by explaining the relevant factors for monitoring socioeconomic indicators through high-level concepts using the TCAV method that can be applied to various geographies without requiring concept labels specific to the study region.

3. Methods

In this section, we detail our explainability pipeline presented in Figure 1 for post-hoc concept-based explanations of socioeconomic outcomes. These methods usually identify concepts as directions in the latent space of an already trained model. This makes their interpretation challenging, as the post-hoc analysis might yield redundant and ambiguous concepts [8]. To overcome this, we rely on a contrastive pretraining technique to enforce continuous ordering of the latent space embeddings based on the target outcome. Such a representation can improve the interpretability of the tested concepts, as it forces them to align along intervals of the target outcome, In summary, the presented approach consists of three steps. Firstly, we train the feature encoder using Rank-N-Contrast (RNC) loss, which optimizes the feature extractor to produce embeddings that are continuously ordered according to the target outcome. Next, we probe a linear layer on the frozen embedding space to regress an output score. Lastly, we analyze the learned embedding through Testing using Concept Activation Vectors (TCAV), a post-hoc concept-based interpretability method.

3.1. Rank-N-Contrast

Rank-N-Contrast (RNC) is a contrastive feature learning method for regression that uses the proximity in label values to form positive and negative pairs within a given batch. In this study we utilize it to pre-train the feature extractor to provide feature embeddings that are aligned with the target regression scores, thereby guiding the expressivity of features in the latent space towards only being relevant for the regression task. Specifically, RNC *"ranks the samples according to their target distances, and then contrasts them against each other based on their relative rankings"* [32]. Consider an anchor image $\mathbf{v}_i$ within a batch of images. For this anchor and a pairing candidate $\mathbf{v}_j$, it samples pairs which satisfy the condition $S_{i,j} := \{v_k \mid k \neq i, d(\tilde{y}_i, \tilde{y}_k) \geq d(\tilde{y}_i, \tilde{y}_j)\}$, where $d(\cdot,\cdot)$ is a label distance metric, while $\tilde{y}_i$, $\tilde{y}_j$ and $\tilde{y}_k$ are the labels of the i-th, j-th and k-th image in the batch, respectively. Subsequently, the feature vectors of all sample pairs are compared using the average negative log-likelihood over all pairings in the batch. By iterating over each sample as an anchor, we can calculate the following loss for the entire batch:

$$\mathcal{L}_{RNC} = \frac{1}{2N}\sum_{i=1}^{2N}\frac{1}{2N-1}\sum_{j=1,j\neq i}^{2N} -log(\sigma(v_i, v_j, S_{i,j}), \tag{1}$$

where the Softmax function σ is given as follows:

$$\sigma(v_i, v_j, S_{i,j}) = -\log\frac{\exp(\text{sim}(v_i, v_j)/\tau)}{\sum_{k\in S_{i,j}}\exp(\text{sim}(v_i, v_k)/\tau)}, \tag{2}$$

where τ is a temperature parameter, and $\text{sim}(\cdot,\cdot)$ measures the similarity between the feature embeddings for a pair of images in the batch. Minimizing this loss enforces the feature encoder to produce similar embeddings for images with similar label values, thus ordering the image embeddings in the latent space according to their labels. After this pretraining step, a simple linear layer can be added to the model and trained on the frozen feature embeddings to regress the actual regression labels.

3.2. Testing with Concept Activation Vectors

We utilize *Testing with Concept Activation Vectors* (TCAV) [17] as an approach for testing the relatedness of concepts to socioeconomic outcomes. TCAV allows uncovering meaningful directions in the latent space of a model based on a previously defined concept dataset and can measure the model's sensitivity to these concepts. Each concept is represented by a set of images sharing similar visual characteristics. These images are usually derived from external probe dataset [24] or automatically learned from the dataset [12]. The workflow of TCAV can be divided into two main steps: 1) learning Concept Activation Vectors (CAVs) in the latent space of the trained model, and 2) testing the sensitivity of a concept to the model predictions.

3.2.1 Learning Concept Activation Vectors

A concept activation vector (CAV) represents a direction in the latent space where the examples of a concept are located. To learn a CAV in a hidden layer of the model, first, the datasets of images for the target concept, and for a random concept are fed into the trained deep learning model, and their activations from the target hidden layer are stored. These activations are used as input to a linear classifier that

learns to discriminate the activations of images representing the concept against the activations of the images of the random concept. The normal vector of the hyperplane is interpreted as the CAV of the respective concept. CAVs can be trained in hidden layers of various depths of the model. This approach relies on the assumption of the linear separability of concepts in the latent space which is supported by several works [4, 6] which have shown that linear feature disentanglement can often be observed in the hidden layers.

3.2.2 Concept sensitivity testing

To evaluate the influence of a concept on the model outcomes, the conceptual sensitivity $S_{k,l}$ of a CAV v_l in layer l to the model output k is determined. It is computed as the directional derivative of the output logit $h_{l,k}$ to v_l with the following equation:

$$S_{k,l} = lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(x) + \epsilon v_l) - h_{l,k}(f_l(x))}{\epsilon} = \nabla h_{l,k} * v_l \tag{3}$$

where $f_l(x)$ are the activations for the input x at the layer l of the deep neural network. Thus, the sensitivity score can be related to change in the model prediction when an instance is perturbed in the direction of a concept. Next, the individual sensitivities are then aggregated to form the overall TCAV score. It is defined as the percentage of images that show a positive sensitivity to the CAV.

$$TCAV_Q = \frac{|x \in X_k : S_{k,l}(x) > 0|}{|X_k|} \tag{4}$$

One caveat when computing the sensitivity scores is the calculation of the gradients. They capture first-order derivatives and thus may lead the attribution map to focus on irrelevant features [27]. To circumvent this issue, we used attribution maps generated via *Integrated Gradients* (IG) that can reduce the problematic locality of plain gradients. They are computed by aggregating the gradients of the model predictions for instances that lie along a straight line on the path from baseline image x' to the input image x with the following equation:

$$\text{IG}(x) = (x - x') \int_0^1 \frac{\partial h_{l,k}(f_l(x' + \alpha(x - x')))}{\partial x} d\alpha \tag{5}$$

For images, the baseline is usually set to a black image, which marks the absence of image features.

4. Experimental Setup

4.1. Reference data

We conduct our experiments on two tasks posing the problem of estimating socioeconomic indicators from aerial imagery. Firstly, we use a dataset that consists of aerial images paired with a winsorized average household income estimated from the tax return for 5 major cities in France in 2015 [1]. The images have a resolution of 20 cm and are of size 1000 x 1000. In line with the authors' workflow of training individual models for cities, we trained one model for Paris, which resulted in 18 446 pairs of aerial imagery and income labels. Similar to [1], we first partition the instances according to the 5 quantiles of their income distribution. Subsequently, we perform stratified random sampling of 64 % of the data into training, and 16 % and 20% in validation and test sets, respectively. For our second task, we use a dataset of liveability reference scores over the Netherlands, which combines the livability values derived from the Leefbaarometer project [19] with aerial images made publicly available by the Dutch government. The resulting dataset covers 13 Dutch cities and includes 51.781 labeled image patches at a 1-meter resolution and a size of 500 x 500. We use the geographically stratified train/val/test splits made available by the authors [21].

4.2. Concept data

We rely on land-cover classes to define visual concepts representing natural and urban areas that can be used to explain the model workings in socioeconomic studies. Concretely, we subset land cover annotations from the FLAIR dataset [11]. It contains aerial imagery covering France at a spatial resolution of 20cm/pixel, as well as land cover class masks. In total, the labels differentiate between 19 different land cover classes, and in our experiments, we extract the following 7 visual concepts: *Water, Natural Vegetation, Agriculture, Impervious Surface, Sparse Residential, Medium Residential*, and *Dense Residential*. The first four concepts are simple visual primitives encoding mostly color and texture features. They are extracted from patches having $> 90\%$ coverage of the following four subsets of land cover classes: *Water, Natural Vegetation, Agriculture* and *Impervious Surface*, respectively. The other three concepts represent more complex visual primitives that were extracted by aggregating land cover patches into three Local Climate Zone [26] classes based on the amount of built-up land visible in each image: *Dense Residential ($> 90\%$ Buildings), Medium Residential (40 to 60% Buildings, 40-60% Vegetation or Agriculture)*, and *Sparse Residential (10-30% Buildings, 70-90% Vegetation or Agriculture)*. Visual examples of our classes can be found in the appendix, in section 8. While we use a set of pre-defined concepts in this research, it should be noted that concepts in remote sensing have stricter conditions on when they may be used. We discuss these conditions in more detail in section 5.3.

4.3. Inference and Model Training

We predict the socioeconomic outcomes using a Resnet-50 encoder [15] for feature extraction whose output is fed to a linear layer that regresses the socioeconomic outcome.

To train the model, we followed a similar training procedure as in [32]: First, we optimize the encoder for 400 iterations to minimize the RNC loss (the hyperparameter details are provided in Section 7, Appendix). Next, we freeze the weights of the feature encoder and train a single linear layer that uses the encoder output to predict the socioeconomic outcomes. The weights of the linear layer were optimized for 100 epochs with the $\mathcal{L}_1$ loss and we selected the model yielding the best R^2 score on the validation set. Our code is provided at `github.com/IvicaObadic/rnc-4-visual-concept-explanations`.

5. Results

5.1. Prediction of Socioeconomic Indicators

Table 1 shows the effect of Rank-N-Contrast pretraining on the prediction results. It illustrates that the task-specific pretraining leads to a better fit for the household income dataset than the baseline model trained with an L_1 loss. Concretely, the R^2 improves for the contrastive pre-trained encoder by around 0.1 for both evaluation sets, validation, and test. Similarly, Kendall's τ correlation improves by around 0.08 on both sets. Regarding the liveability dataset, the contrastive pretraining procedure yields results comparable results to the encoder directly trained for predicting livability scores with the L_1 loss.

Table 1. **Impact of pretraining with Rank-N-Contrast (RNC) loss on prediction results. It results in a better prediction performance for the income predictions and comparable results for liveability.**

		val set		test set	
dataset	**objective**	R^2	τ	R^2	τ
Income	$\mathcal{L}_1$ loss	0.53	0.54	0.55	0.56
	RNC ($\mathcal{L}_1$)	0.62	0.61	0.65	0.64
Liveability	$\mathcal{L}_1$ loss	0.7	0.64	0.49	0.53
	RNC ($\mathcal{L}_1$)	0.68	0.66	0.49	0.53

To better understand the effect of task-specific pretraining on the latent space, in Figure 2 we use the t-distributed Stochastic Neighbor Embedding (t-SNE) method [28] to visualize the instance activations after the average pooling layer of the ResNet-50 encoder. The plots in the first column illustrate that the contrastive pre-trained encoder increasingly orders the instances in the latent space in an approximately 1D manifold according to their target label for both datasets. In contrast, the plots visualized in the second column depict that the models optimized to directly predict the regression score seem to result in a more scattered embedding space that fails to capture the continuous nature of the data, as there are multiple clusters of instances sharing similar target values located in various subspaces of the latent space. Finally, we note that for the task-specific pre-trained encoder, the liveability values typically display higher variance in their score compared to the income values for a local area in the embedding space. This can be attributed to the difference in complexity between these two tasks. The liveability reference data is based on 100 factors spanning 5 domains, from which liveability scores are derived [21]. As it is a composite score, the ranking difference between pairs will depend on more complex patterns, which might explain the comparable results of the contrastive pretraining with the supervised trained encoder on the liveability task.

5.2. Concept Explanations

Learning Concept Activation Vectors As described in Section 3.2, an essential step in TCAV is learning a concept activation vector (CAV) that encodes the direction in the latent space where the examples of a concept are located. In our study, we learn the CAVs per layer in the latent space of the ResNet-50 encoder by training a one-vs-all linear SVM classifier to discriminate the examples of the concept from the examples of all other concepts based on their layer activations. Figure 3 shows that while the concepts are already mostly linearly separable in the average pooling layer of the baseline encoder trained with the L_1 loss, enforcing a continuously ordered latent space with the Rank-N-Contrast pre-trained encoder enables close-to-perfect discrimination, with substantial improvements in the medium residential concept for the liveability dataset and impervious surfaces, vegetation, agriculture, and sparse residential concepts in both datasets.

Alligning Concepts to Socioeconomic Outcomes Not imposing any constraints on the latent space during model training is challenging for post-hoc concept interpretability as the resulting model might yield ambiguous CAVs that are hard to interpret [8]. As shown in the previous section, the Rank-N-Contrast pretraining procedure ensures that the embeddings lie on a manifold continuously ordered according to the target value. This impacts the learned CAVs as the concept examples are also projected along this manifold. Further, the high linear separability of the concepts on this manifold shown in Figure 3 suggests that the CAVs point to distinctive ranges of socioeconomic outcomes. We verify this by performing a similar analysis to [17] that identifies the most similar concept to each instance in the socioeconomic datasets based on the cosine similarity of the instance embeddings and the CAVs. The graphs on the left-hand side of Figure 4 reveal that in the contrastive pre-trained encoder, the concepts are clustered on typically non-overlapping intervals of the target values. In detail, for household income, the lower income areas are associated with the natural concepts of vegetation and water. Next, middle and high-income areas are aligned with sparse residential areas and agriculture fields while those with the highest income

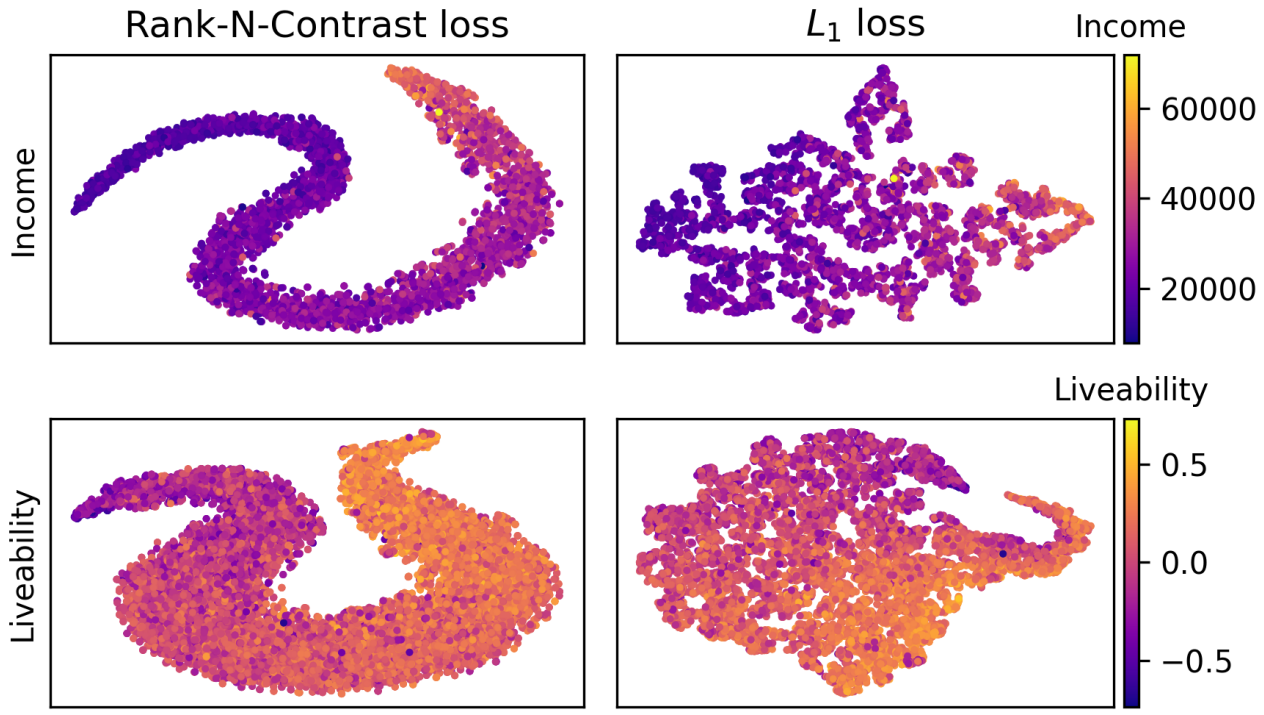

Figure 2. **Household income instance activations in the average pooling layer visualized with t-SNE.** By pre-training with Rank-n Contrast, the latent space can be ordered according to the task regression values, rather than the visual features of images. This results in an embedding space that aligns with the socioeconomic outcomes, and therefore better suited for interpretability.

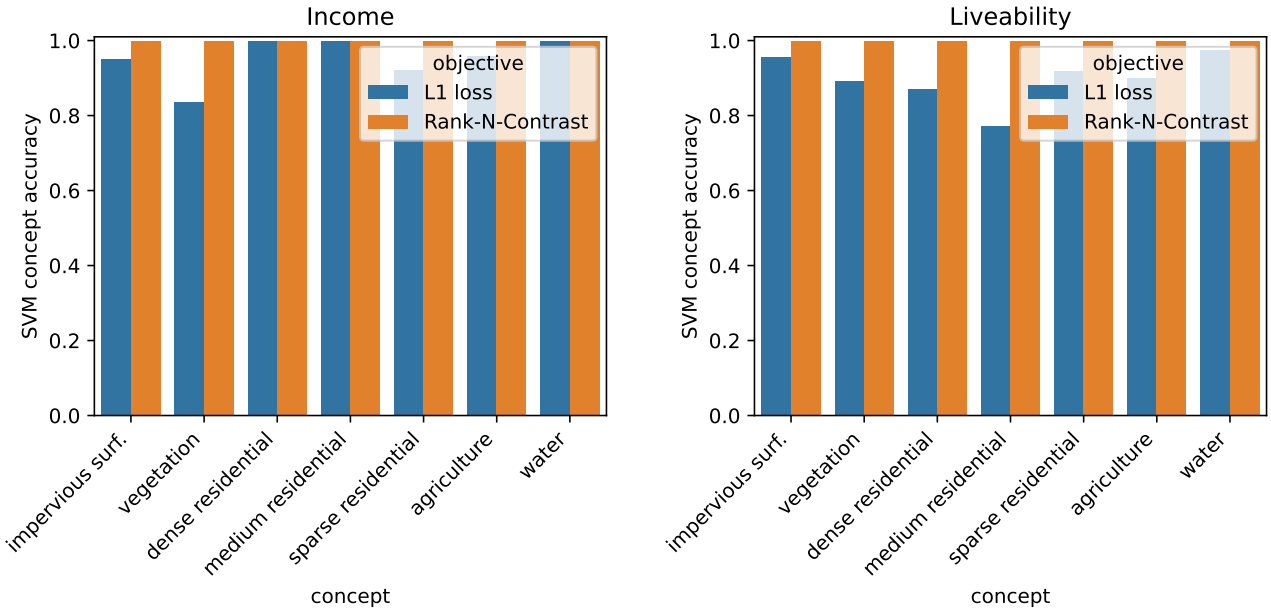

Figure 3. **Concept accuracy in the average pooling layer of the Resnet-50 encoder for the household income (left) and livability (right) datasets.** The contrastive pretraining improves the linear separability of the concepts in the latent space.

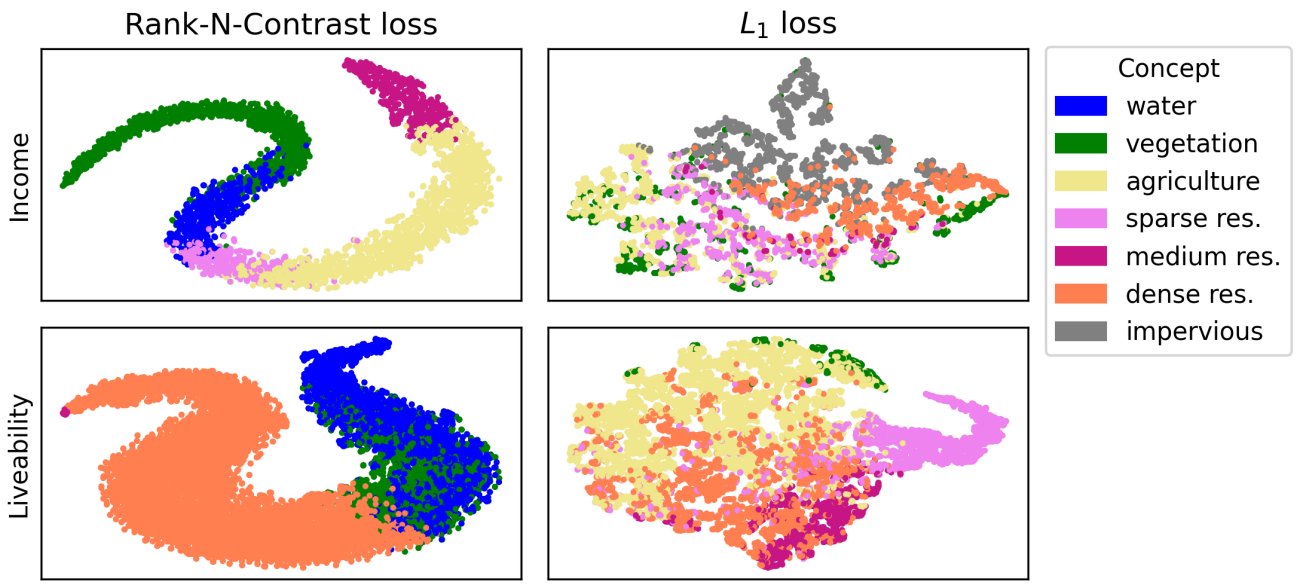

Figure 4. **Instance to concept alignment as seen in the embeddings of each model.** The instances are colored according to the concept with the highest cosine similarity (after normalizing the similarities per concept with the L_2 norm). The most similar concept to an instance can be interpreted as the concept that is most closely aligned to the instance's socioeconomic outcome.

resemble medium residential areas. Regarding liveability, dense residential areas characterize the highest portion of the liveability range, particularly for the low and medium-liveable areas. Subsequently, the highest liveable areas are closely aligned with the natural concepts of water and vegetation. In summary, these results show that the Rank-N-contrast pretraining procedure improves the interpretability of the concept explanations as it enables one to intuitively associate the concepts to a continuous interval of values for the socioeconomic indicators. Conversely, the plots on the right-hand side of Figure 4 visualize that for both datasets, the models trained with the L_1 loss offer a limited understanding of how the concepts relate to the socioeconomic indicators as they typically do not form clear clusters in the latent space.

Insigths into Urban Planning with TCAV Measuring the sensitivity of the concept to the model prediction with Equation (3) of Section 3.2.2 enables us to understand how the prediction changes when the visual patterns of the instances are perturbed in the direction of the concept. Furthermore, the concept clusters resulting from Rank-N-Contrast pretraining can unveil the correlation for intervals of socioeconomic indicators with the different urban and natural concepts. To shed light on the potential of such analysis for urban studies, in Figure 5 we visualize the magnitude of the TCAV values for the concept of vegetation. The left plot shows that perturbing the areas with low income in the direction of vegetation increases the model prediction for income. This positive effect decreases for areas with higher income. In other words, low-income areas can be seen rising in income as vegetation increases, with diminishing returns. The lowest-income areas in Paris can be found in the north and the northeast of the city, which is less green than most higher-income areas. For liveability, the opposite pattern is observed, where adding natural vegetation and water is only beneficial to the most liveable areas. While this may seem counterintuitive at first, it is important to consider that the liveability score is a composite socioeconomic outcome, comprising multiple dimensions at once. Neighborhoods in more densely populated areas often have better amenities and higher income opportunities. However, what is often lacking in these areas is access to nature. Conversely, while rural non-urban areas are abundant in natural vegetation, they are known to have lower incomes and access to amenities. The sensitivity plots of other scores can be found in Appendix, Section 9.

5.3. Discussion - Defining Remote Sensing Concepts

The use of concepts as a means of interpretability in natural images is well-studied, with many datasets available that contain concept classes. However, given the more complex and variable nature of remote sensing imagery, the use of concepts and their comparison between studies will strongly depend on definitions, datasets, and sensor resolution used. Firstly, as the resolution of an image decreases, individual objects lose their shape, and only geographical features, which are "an abstraction of a real-world phe-

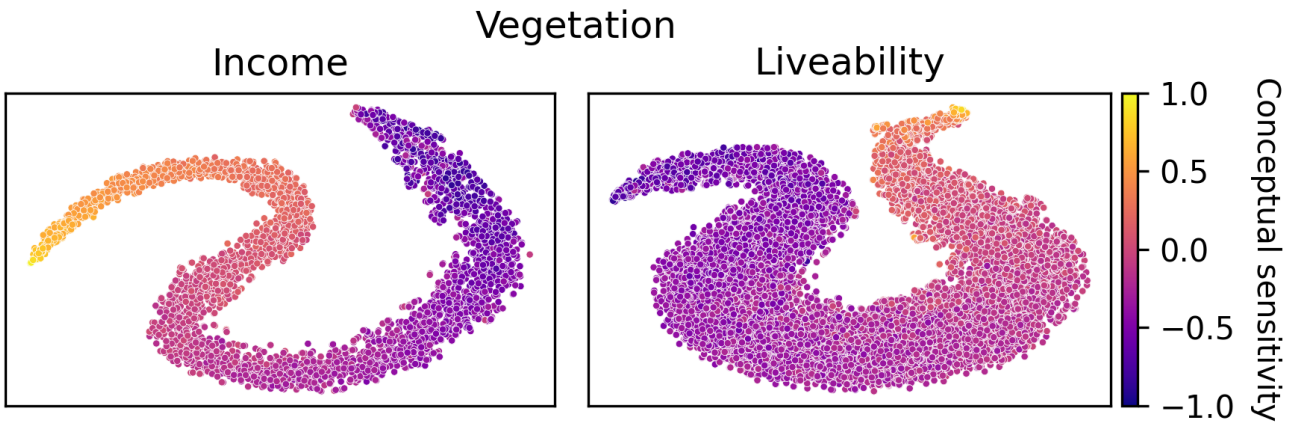

Figure 5. **The TCAV sensitivity of the vegetation concept for the income (left) and liveability (right) datasets.** The magnitude values are normalized in the range [-1, 1] with min-max normalization. For income, the sensitivity to vegetation is highest among the lowest income areas and it diminishes for the areas with higher income. In other words, adding vegetation to low-income areas increases the perceived income of the neighborhood by the model. For liveability, we observe the opposite effect, as the strongest increase in perceived liveability by the model in highly liveable areas can be achieved by increasing their amount of natural vegetation.

nomenon" [16], remain. As a result, the definition of a geographical concept varies with the spatial scale that it is observed on [30]. Concept-centric approaches for remote sensing therefore need to contend with variable concept definitions, making large-scale remote sensing concept datasets harder to manage. Beyond concept definition uncertainties, common geographical products such as land cover maps may contain classes that are not visible with optical sensors [9]. As Arvor et al [5] note, geographic concepts suffer from vagueness in their definitions. Often, concepts are defined using qualitative metrics. For instance, "high NDVI" may define a forest. However, values ascribed to qualitative descriptors such as "high" may differ between studies. A second-factor affecting uniform concept definitions is determined by the ambiguity of a concept, where it has "no crisp boundaries and may shift its meaning over time as new technologies appear, as people develop new habits, and in general, as the social and physical context of the term evolves" [29]. Finally, the cultural context of a term makes defining geographical concepts difficult, as these concepts may readily vary between cultures [22]. As concept-based explanations continue to mature, such considerations must be taken into account.

6. Conclusion

In this paper, we presented a pipeline that enables intuitive concept explanations of the deep learning models for socioeconomic outcome prediction. It consists of task-specific contrastive pretraining with the Rank-N-Contrast method and the TCAV approach for post-hoc concept testing. Our results show that the use of Rank-N-Contrast projects latent space embeddings in a continuously ordered manifold based on socioeconomic outcomes. This resulted in improved model performance for income prediction in Paris and yielded comparable results for liveability predictions in the Netherlands. The learned representation also contributed to intuitive concept explanations as we have shown that the concept activation vectors form clusters along continuous intervals of socioeconomic outcomes. For example, our analysis revealed that dense residential areas are typically associated with low liveability while more liveable areas are those that are close to the natural concepts of water and vegetation. The ordering of the concepts in the latent space is also beneficial for understanding their sensitivity to the model predictions. We illustrated this in Figure 5 which indicates that perturbing the instances of low income towards the vegetation concept improves their income score whereas for liveability this effect was observed for the highly liveable areas. Such insights open up the possibilities of our method for new insights into urban studies. Finally, we showed that our approach can be applied in different geographical locations as unlike concept bottleneck models, it does not require a set of concept labels specific to the location of the target study available during training.

Acknowledgments This work is jointly supported by the Munich Center for Machine Learning and by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001). We thank Dino Ienco, Nassim Ait Ali Braham, and Andrés Camero for the fruitfull discussions.

References

[1] Jacob Levy Abitbol and Marton Karsai. Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nature Machine Intelligence*, 2(11):684–692, 2020. 1, 2, 4

[2] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023. 1

[3] Julius Adebayo, Justin Gilmer, Ian J. Goodfellow, and Been Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. *CoRR*, abs/1810.03307, 2018. 1, 3

[4] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. 4

[5] Damien Arvor, Mariana Belgiu, Zoe Falomir, Isabelle Mougenot, and Laurent Durieux. Ontologies to interpret remote sensing images: why do we need them? *GIScience & Remote Sensing*, 56(6):911–939, 2019. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15481603.2019.1587890. 8

[6] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations, 2017. 4

[7] Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021. 1

[8] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020. 3, 5

[9] Russell G. Congalton, Jianyu Gu, Kamini Yadav, Prasad Thenkabail, and Mutlu Ozdogan. Global Land Cover Mapping: A Review and Uncertainty Analysis. *Remote Sensing*, 6(12):12070–12093, 2014. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute. 8

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 1

[11] Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sébastien Giordano, et al. Flair: a country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[12] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019. 1, 3

[13] Mara Graziani, Vincent Andrearczyk, Stéphane Marchand-Maillet, and Henning Müller. Concept attribution: Explaining cnn decisions to physicians. *Computers in biology and medicine*, 123:103865, 2020. 1

[14] Ola Hall, Mattias Ohlsson, and Thorsteinn Rögnvaldsson. A review of explainable ai in the satellite data, deep machine learning, and human poverty domain. *Patterns*, 3(10), 2022. 1, 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[16] ISO. ISO19101 "Geographic Information - Reference Model.". https://www.iso.org/standard/59164.html, 2014. [Online; accessed 14-April-2024]. 8

[17] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, pages 2668–2677. PMLR, 2018. 1, 2, 3, 5

[18] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. "help me help the AI": Understanding how explainability can support human-AI interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 2023. 1

[19] Kees Leidelmeijer, Gerard Marlet, Roderik Ponds, René Schulenberg, and Clemens van Woerkens. Leefbaarometer 2.0: Instrumentenontwikkeling. 1, 4

[20] Alex Levering, Diego Marcos, and Devis Tuia. On the relation between landscape beauty and land cover: A case study in the uk at sentinel-2 resolution with interpretable ai. *ISPRS journal of Photogrammetry and Remote Sensing*, 177:194–203, 2021. 1, 2

[21] Alex Levering, Diego Marcos, Jasper van Vliet, and Devis Tuia. Predicting the liveability of dutch cities with aerial images and semantic intermediate concepts. *Remote Sensing of Environment*, 287:113454, 2023. 1, 2, 4, 5

[22] David M. Mark. Toward a theoretical framework for geographic entity types. In *Spatial Information Theory A Theoretical Basis for GIS*, pages 270–283, Berlin, Heidelberg, 1993. Springer. 8

[23] Ye Ni, Xutao Li, Yunming Ye, Yan Li, Chunshan Li, and Dianhui Chu. An investigation on deep learning approaches to combining nighttime and daytime satellite imagery for poverty prediction. *IEEE Geoscience and Remote Sensing Letters*, 18(9):1545–1549, 2020. 1

[24] Vikram V Ramaswamy, Sunnie SY Kim, Ruth Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10932–10941, 2023. 3

[25] Sanja Scepanovic, Sagar Joglekar, Stephen Law, and Daniele Quercia. Jane jacobs in the sky: predicting urban vitality with open satellite data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–25, 2021. 1, 2

[26] I.D. Stewart and T. Oke. Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93:1879–1900, 2012. 4

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. 4

[28] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 5

[29] R. J. Wieringa. *Design Methods for Reactive Systems: Yourdon, Statemate, and the UML.* Elsevier LTD, Oxford, Boston, 2002. 8

[30] Curtis E. Woodcock and Alan H. Strahler. The factor of scale in remote sensing. *Remote Sensing of Environment*, 21(3): 311–332, 1987. 8

[31] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):2583, 2020. 1

[32] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 5, 1

[33] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 1

[34] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2131–2145, 2019. 1