ARTICLES
https://doi.org/10.1038/s42256-020-00243-5
nature machine intelligence

Interpretable socioeconomic status inference from aerial imagery through urban patterns

Jacob Levy Abitbol[1]✉ and Márton Karsai[1,2]✉

Urbanization is a great challenge for modern societies, promising better access to economic opportunities, but widening socioeconomic inequalities. Accurately tracking this process as it unfolds has been challenging for traditional data collection methods, but remote sensing information offers an alternative way to gather a more complete view of these societal changes. By feeding neural networks with satellite images, the socioeconomic information associated with that area can be recovered. However, these models lack the ability to explain how visual features contained in a sample trigger a given prediction. Here, we close this gap by predicting socioeconomic status across France from aerial images and interpreting class activation mappings in terms of urban topology. We show that trained models disregard the spatial correlations existing between urban class and socioeconomic status to derive their predictions. These results pave the way to build more interpretable models, which may help to better track and understand urbanization and its consequences.

Cities have become the economic bedrock of modern nations. In little more than a century, they have gone from concentrating 13% to an estimated 55% of the world population, with 600 of them currently accounting for around half of the global economic output (US$34 trillion of gross domestic product)[1]. This trend will probably accelerate in the coming years, as an estimated three billion people will move into cities by 2030. This increased urbanization is in turn a key driver of development, as cities provide a national platform for shared prosperity. Indeed, the concentration of people in cities generates agglomeration economies, where the sheer population density facilitates the moving of goods, people and ideas by removing the physical spaces between people and firms, thus increasing the returns of urban proximity[2].

Nevertheless, while urbanization can drive economic dynamism and social development, it can also create enormous social challenges. The management of natural hazards and pollution, the exclusion of the poor from a city's socioeconomic fabric and the subsequent surge of social and economic inequalities have become some of the pressing issues that a modern metropolis needs to address. This last issue is especially acute in some cities with the greatest concentration of wealth. For example, Gourevitch et al. reported a 24-year difference in the average lifespan of citizens living in neighbouring census tracts in San Francisco[3]. Meanwhile, greater Paris contains the second and fifth of the top five poorest and richest census tracts, respectively, in all of France[4]. Providing a solution to these challenges is of paramount importance to fulfil the economic and social promises that cities hold and to avoid them becoming sources of social and political instability.

The successful development and deployment of urban solutions to address these issues requires both spatially fine-grained socioeconomic information as well as a detailed understanding of how wealth and the underlying urban topology are entangled. Although this question has been thoroughly addressed before in previous urban sustainability studies[5–8], the socioeconomic maps used were generally coarse-grained (roughly in the order of km^2), the finest level of description being of course that provided by the census, thus hindering any detailed analysis. To cope with this, other works have relied on socioeconomic proxies, mostly derived from large-scale digital datasets, to propose accurate and highly detailed socioeconomic status (SES) estimations. These include communication patterns in call detail records[9] and social media[10] or even restaurant data[11].

Most notably, some works have relied heavily on large collections of satellite and street imagery to train deep-learning models to predict wealth either from visual features[12] or by predicting features known to correlate with wealth (such as night-time light intensity[13] or models of cars in census tracts[14]). This last approach has been further refined by works aimed at making interpretability a key point of study when predicting the SES of an area. For example, Ayush et al. sought correlations between high-level object detections and poverty scores in Uganda and found the presence of trucks to be predictive of local scale poverty[15]. On the other hand, Sheehan et al. relied on information on economic development of Wikipedia articles of a given country to predict poverty and observed that articles containing words like 'school', 'university' or 'hospital' tended to be linked to wealthier places[16]. By using features that are explicitly linked to wealth, these works missed explaining the outputs of more classical deep-learning models, which are trained end-to-end to predict wealth.

Surprisingly, despite the overall use and reliability of these last methods, no previous work has looked consistently at the features learned by these models, nor tried to uncover any existing correlations between the existing urban topology of a city and the high-resolution socioeconomic map the model is tasked with predicting.

Our aim is to close this gap by training a deep-learning model to predict the SES of a given location from its aerial image and in turn interpret its activation maps in terms of the underlying urban topology. More precisely, we first generate a complete description of five French cities in terms of socioeconomic and land-use data, as well as aerial imagery. Subsequently, after merging the aerial imagery with the corresponding socioeconomic maps, we train a convolutional neural network (CNN) model to accurately predict the SES of inhabited tiles. Next, by relying on a gradient-weighted

[1]University of Lyon, Inria, CNRS, ENS de Lyon, Université Claude Bernard Lyon 1, Lyon, France. [2]Department of Network and Data Science, Central European University, Vienna, Austria. ✉e-mail: jacobolevyabitbol64@gmail.com; karsaim@ceu.edu

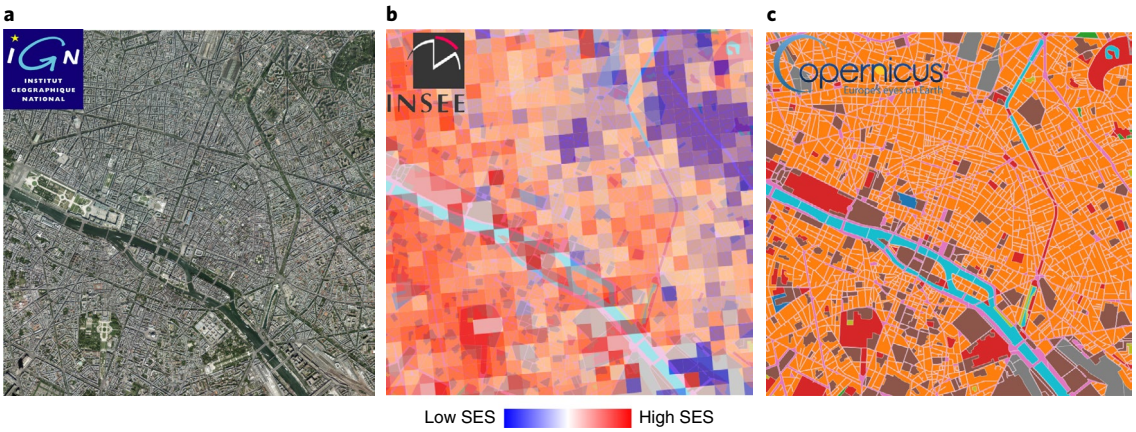


Fig. 1 | Sample of overlaid datasets (Paris). a, A 5 km × 5 km aerial tile (20 cm per pixel). **b**, Spatial distribution of income: each patch corresponds to a single 200 m × 200 m area with precise income data. **c**, Land cover map of the same area, where each colour represents a different urban class (not shown here for clarity). Sample sizes for predicted and true values per socioeconomic and urban class are provided in Supplementary Tables 2 and 3.

Table 1 | Dataset information

Dataset	Collection timespan	Spatial resolution	Coverage
Socioeconomic census	2015	200 m × 200 m cells	Metropolitan France
Aerial imagery	2013–2016	20 cm per pixel	Metropolitan France
Urban atlas	2012	Maps accurate to 5 m	EU cities of more than 100k inhabitants

class activation mapping (Grad-CAM) for computing attribution maps[17–19], we generate high-resolution class-discriminative activation maps, which are projected back onto the original image and overlaid with land-use data. This framework enables the inference of SES at scales rarely seen before, while also indicating precisely the features contained in the actual urban environment used for prediction by the model. In doing so, we find that, when inferring SES, our CNN models disregard existing correlations between land-use and socioeconomic data and focus mostly on features contained within residential areas.

Results

Aerial, socioeconomic and land-use datasets and their combination. The fine-grained prediction of SES from remotely sensed images requires both a detailed socioeconomic map of the area of interest and a dataset of aerial images provided at a sufficiently high resolution for the CNN to be able to detect localized features. Throughout this study, we rely on high-resolution aerial imagery taken between 2013 and 2016 of the complete French metropolitan territory (at 20 cm per pixel), issued by the National Geographical Information Institute (IGN)[20]. This orthophotography dataset is distributed as a series of georeferenced 5 km × 5 km tiles provided at the department (administrative delimitation) level, as shown in Fig. 1a. This collection is then complemented with socioeconomic data shared by the French National Institute of Statistics and Economic Studies (INSEE) in 2019[21]. This corpus describes the winsorized average household income, estimated from 2015 tax returns in France, for each four-hectare (200 m × 200 m) square patch (a 'census cell') across the whole French territory (Fig. 1b). A summary of the main features of each dataset is provided in Table 1. The combination of the socioeconomic map and the aerial data requires us to identify the aerial image (an 'aerial tile') corresponding to each 200 m × 200 m census cell, therefore enabling us to obtain, for every socioeconomic patch, its corresponding aerial view. For details regarding the winsorization of the census data or its spatial matching with aerial images, see Supplementary Section 1.

Next, to study the correlations between the distribution of wealth and the urban patterns, we gather land-use data shared by the European Environment Agency through the 2012 European Union Urban Atlas project. This dataset provides high-resolution land cover maps of roughly 700 urban areas of more than 100,000 inhabitants in EU28 and European Free Trade Association (EFTA) countries. It does so by segmenting each city into detailed polygons categorized into one of 27 standardized land-use classes. In our study, only 19 classes were used, as infrequent and similar classes were respectively discarded and merged. Further information about the generation of this dataset is provided in the EU Urban Atlas[22] and explained in the Methods.

Following the spatial combination of these datasets, we single out five major cities located within the French metropolitan territory for further analysis. These were Paris, Lyon, Marseille, Nice and Lille. Geographical boundaries for each city/urban area are obtained from the ensuing land use data. For each city, income values from the socioeconomic patches are partitioned into one of five ($n_{\mathrm{SES}} = 5$) socioeconomic classes defined by the five quantiles of the city-wise income distribution, so that classes one and five correspond respectively to the bottom and top 20% of earners in each city. In doing so, we obtain the spatial distribution of income specific to each city, as is shown in Fig. 3a for Paris, where each coloured pixel corresponds to a single socioeconomic patch. Similar results for the other cities can be found in Supplementary Figs. 2–5 together with a detailed description of the used datasets. Note that all the collected datasets can be openly accessed and emanate from a single four-year time window hence minimizing their temporal misalignment.

Inferring SES in cities. To estimate the SES of a given location, we modify and train an EfficientNetB0[23] model on socioeconomic and aerial tile pairs to predict the former from the latter. The trained model takes individual aerial tiles $s \in S$ set of tiles as input and predicts an SES label $\hat{y}(s) \in [1, \ldots, n_{\mathrm{SES}}]$ for each of them. This model was chosen over more classical models (like VGGs or ResNets) mostly for its architectural design: by making use of an effective

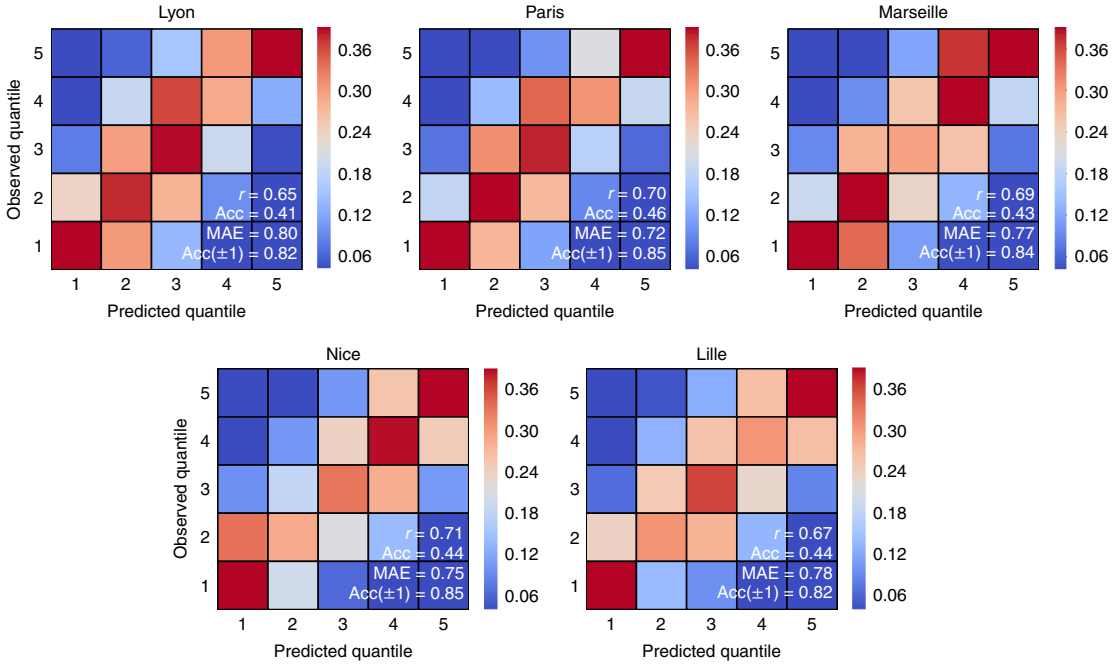


Fig. 2 | Observed performance of models trained to predict wealth in French cities: confusion matrices between predicted and observed SES classes. In each plot, the Pearson correlation coefficient (r, $P < 0.01$) as well as accuracy (Acc), mean average error (MAE) and accuracy within ±1 (Acc(±1)) classes are provided. All results are averaged over the k cross-validation folds (with $k = 5$).

compound coefficient to scale up CNNs, EfficientNets have been shown to achieve better accuracy and efficiency than other convolutional neural networks, as also proved to be the case in our setting (for a comparison to other models see Supplementary Table 1). This enabled us to use the original fine-grained resolution of the aerial imagery as input to predict the SES at a lower computational cost. The architecture and parametrization of our model are further described in the Methods.

The modified EfficientNetB0 models are separately trained on each city to predict five-class SES from 200 m × 200 m aerial tiles. To assess the quality of the predictions, we computed the metrics proposed by Suel et al.[12] for each city in our dataset. We report the fivefold averaged cross-validation performance by means of separate confusion matrices in Fig. 2.

In each case, perfect prediction would correspond to red diagonal and blue off-diagonal cells, while the greater the deviation from this pattern the greater the classification error made. We also report the Pearson correlation coefficient r, the mean average error (MAE), accuracy (Acc) and accuracy within one class (Acc(±1)) values between the observed and predicted SES classes in the figure keys. Note that Pearson's r is added as a means of comparing our results to the aforementioned work[12], despite both the predicted and true SES classes being discrete variables. All values were computed for samples not seen either during training or validation of the model.

The quality of the predictions tends to be quite similar between cities, with the average Pearson correlation between true and predicted quantiles varying between 0.65 for Lyon and 0.71 for Nice. Furthermore, in all cities, the MAE between predicted and observed socioeconomic classes is bounded between 0.72 for Paris and 0.80 for Lyon. Although some misclassifications occur between neighbouring quantiles, all tiles are classified within ±1 of their class with similar accuracies to those reported in previous studies (ranging from 0.82 to 0.85)[12]. Interestingly, the classes predicted by our model with the lowest accuracy tend to vary between cities. For example, in Paris and Lyon, the fourth quantile appears to be the hardest to predict correctly, while the second and third seem to be most error-prone in Nice and Marseille, respectively.

To visualize our predictions clearly, the observed and predicted spatial distributions of income for Paris are depicted in Fig. 3 (and for the other cities in Supplementary Figs. 2–5). Contrary to other studies[12,14], here, each map pixel represents a single cell corresponding to a unique tile, a socioeconomic class pair. Consequently, predicted values are not aggregated but individually computed. It is therefore interesting to see that neighbouring tiles of equal SES are overall correctly predicted to be of the same socioeconomic class, hence recovering the small-scale spatial homogeneity of the original income distribution (see centre of Paris, Lyon or southwest of Nice and Marseille). At the same time, sub-urban regions characterized by socioeconomic disparities seem to be recovered. This can be seen in the predicted socioeconomic segregation between downtown Paris and its suburbs (Fig. 3), the west of Marseille or the east of Lille (Supplementary Figs. 3 and 5), which further demonstrates that the trained models are able to recover the complex spatial distributions of income featured in our dataset with remarkable resolution and accuracy. Because the models are trained and tested in the same city, the risk of overfitting is considerable and difficult to assess in the current setting. A detailed discussion of this issue is included in the Supplementary Information (Supplementary Fig. 6).

Grad-CAM-derived model correlations between SES and urban topology. Once all models are trained, we seek to identify the most salient features used by the CNNs to draw their predictions in terms of land use. This is a highly non-trivial task, as it requires mapping the abstract activity patterns of the CNN back onto the original georeferenced space. To solve this, we apply guided gradient-weighted class activation mapping[17] (guided Grad-CAM) to identify visual features underlying the model's predictions. Previous uses of this method include the identification of patterns learned by CNN models trained to predict Alzheimer's disease[18,19] and lung cancer[24], as well as to classify wildlife[25] and plant diseases[26]. Guided Grad-CAM follows the CNN's gradient flow from individual output classes back onto the original image tile to establish an activation map, highlighting the input features most relevant to each class prediction. Figure 4 shows how Guided Grad-CAM (white features on dark

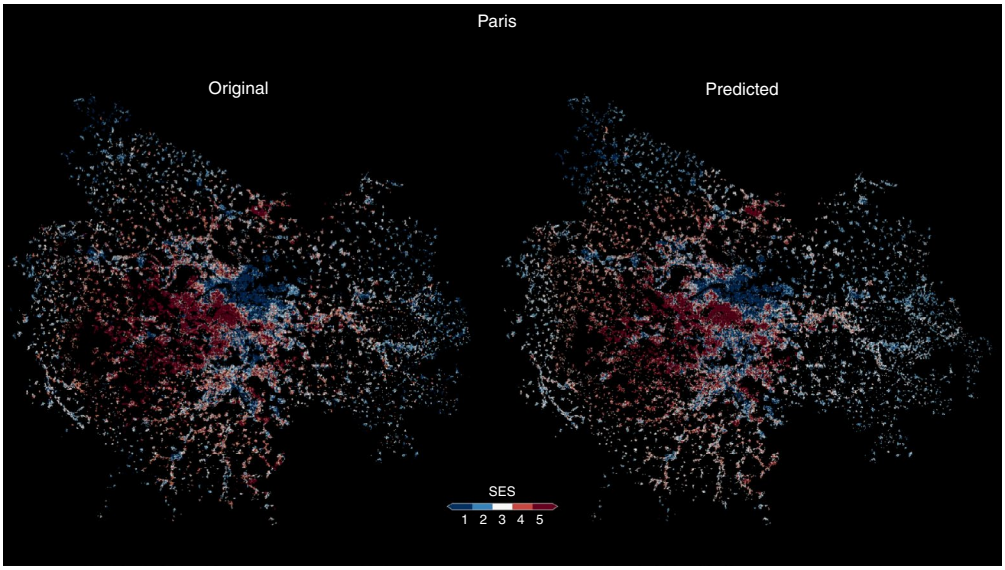


Fig. 3 | Maps of observed and predicted average income for Paris. Each pixel represents a single 200 m × 200 m tile, with colour indicating the average SES of its inhabitants. Maps for the other cities are provided in Supplementary Section 3.

background) is applied to a given aerial tile (Fig. 4a). The overlaid activation maps highlight the visual features that most trigger the prediction for the poorest (Fig. 4b) and richest (Fig. 4c) classes. These activation patterns can then be associated to the corresponding land-use maps shown in Fig. 4d.

In turn, one can compute the overall activation values for different urban class polygons, and see which urban structures are the most salient when predicting the poorest (Fig. 4e) and richest (Fig. 4f) classes. In what follows, we build on this idea by first providing a clear definition on how to measure correlations between saliency maps and urban classes and then probe the model's learned features by evaluating the following two hypotheses:

- H0: All urban classes contain features equally contributing to the model's prediction for the top and bottom SES class.
- H1: Urban class activation maps are pairwise independent; that is, activations for a given urban class are, on average, invariant to activation values for other urban classes.

We treat guided Grad-CAM maps as saliency maps describing pixels that are most determinant to the model's prediction. In doing so, every pixel (i, j) in the saliency map of a given tile $s \in S$ is described both by its guided Grad-CAM activation value A^s_{ij} and the urban class $u \in U$ set of urban classes of the polygon to which it belongs via the mapping f, so that $\forall(i, j), f(i, j) \in U$. Furthermore, we define the total activation for urban class $u \in U$ in tile $s \in S$ to be $A^s_u = \sum_{(i,j)|f(i,j)=u} A_{i,j}$ so that each sample $s \in S$ is characterized by the tuple $(A^s_u)_{u \in U}$ where $A^s_u = \backslash \mathrm{NA}$ for all urban classes $u \in U$ not contained in tile s.

Univariate correlations. To examine the first hypothesis, we introduce the activation ratio for urban class u defined as the ratio between the sum of activations and the expected sum of activations in a random diffusion model concentrated by polygons of urban class u:

$$r^s_u = \frac{A^s_u}{F_u \sum_{k \in U} A^s_k}$$

where F_u is the fraction of area occupied by polygons of class. This metric actually describes how the pixels that are triggered the most for the prediction of a socioeconomic class are distributed across urban classes. For example, if no urban class contains more important features than others (H0), its expected value should be equal to one over all samples and for all urban classes contained in each sample. To obtain an aggregated view on this metric, we also computed the expected activation ratio for urban class u, $x^\sigma_u = \mathbb{E}_\sigma(r^s_u)$ on the set σ of aerial tiles predicted to be of low ($\sigma_{\mathrm{low}} = \{s \in S | \hat{y}_s \in \{1, 2\}\}$) or high ($\sigma_{\mathrm{high}} = \{s \in S | \hat{y}_s \in \{3, 4\}\}$) SES. Furthermore, to provide a more grounded comparison to the actual link between SES and urban patterns, we measured the empirical probabilities $\hat{p}^{\mathrm{low}}_u$ and $\hat{p}^{\mathrm{high}}_u$ that an aerial tile is of low or high SES given that it contains urban class u.

We now identify the salient visual features underlying the model's predictions in terms of land use for samples predicted to be of low and high SES. In this analysis, we refer to an urban class u being over-activated if its expected activation ratio is greater than one, that is, $x^\sigma_u > 1$. Conversely, it is referred to as under-activated whenever its expected activation ratio is below one, that is, $x^\sigma_u < 1$. The results for Paris are shown in Fig. 5, while those of other cities are provided in Supplementary Figs. 7–9 and Extended Data Fig. 1. To ease the description in Fig. 5, we group each urban class into four categories: functional, infrastructure, nature or residential areas. Sample sizes for Fig. 5 are provided in Supplementary Tables 2 and 3.

Looking at the spectrum of activations, certain structural elements tend to be shared across all trained models. Residential areas are over-activated when predicting both low and high SES, with some differences remaining between cities: for models trained to predict SES in Lyon, Lille and Marseille, residential areas tend to activate to a greater extent when predicting high SES, while residential areas appear over-activated in Paris and Nice when predicting low SES.

Infrastructural areas, on the other hand, follow quite a different pattern as they tend to be over-activated for low SES class predictions. For example, railways and motorways are triggered in Lyon and Marseille, while features contained within roads are generally active for all cities.

More generally, analysis of the univariate correlations is evidence that CNN models trained to predict SES in urban areas are mostly reliant on features lying within residential areas to draw their predictions, while the degree to which other amenities are determinant in the prediction shifts from city to city, therefore invalidating H0.

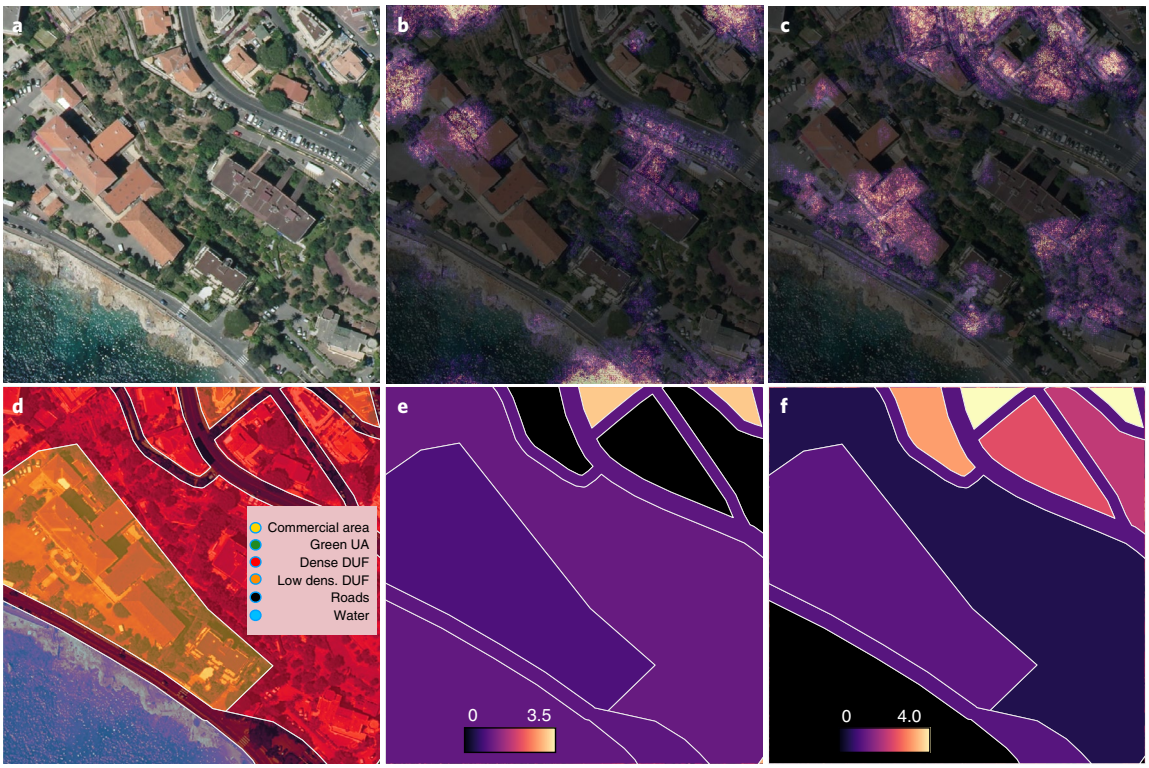


Fig. 4 | Model interpretability studies using guided Grad-CAM. a–c, From an aerial tile (**a**), guided Grad-CAM is used to compute activation maps for the poorest (**b**) and wealthiest (**c**) socioeconomic class. **d–f**, The activation maps are then overlaid with the tile's tessellation into an urban classes polygon (**d**) to compute the normalized ratio of activations per polygon for the poorest (**e**) and wealthiest (**f**) class. UA, urban area; DUF, discontinuous urban fabric. Sample sizes for predicted and true values per socioeconomic and urban class are provided in Supplementary Tables 2 and 3.

Significantly different patterns are observed when looking at the empirical probabilities, $\hat{p}_u^{\text{low}}$ and $\hat{p}_u^{\text{high}}$ for Paris, as shown in Fig. 5b (and for other cities in the Supplementary Information). For example, while the type of residential areas contained in a given tile greatly varies between SES classes, low- and medium-density areas are more prone to being located within high-SES tiles in any city, while very-high-density areas are more likely to be contained within low-SES tiles. Furthermore, interesting patterns can be observed for other categories. For example, tiles containing motorways, railways and commercial/industrial units were mostly associated with low-SES areas, whereas those including natural areas were more likely to be of high SES.

Bivariate correlations. We have so far established that the models we train to estimate SES derive their predictions mostly from features contained within residential areas. Even though the previous analysis sheds light on the features behind the models' prediction for both low and high SES, it is still unclear how the importance of a residential area is affected by the fact it neighbours other types of urban class. To address this question, we mimic the former analysis and measure the expected coactivation ratio for pairs of urban classes u and v, as $c^{\sigma}_{u,v} = \mathbb{E}_{\sigma'}(r^s_u)/\mathbb{E}_{\sigma}(r^s_u)$. Here, $\sigma' = \{s \in \sigma | A^s_u, A^s_v \neq \backslash\,\text{NA}\}$ is the set of aerial tiles where urban class v is present alongside urban class u, thus $\sigma' \subseteq \sigma$. Therefore, for a given urban class u and for any other urban class v, if H1 were to hold, the expected value of the activation ratio $\mathbb{E}_{\sigma'}(r^s_u)$ should remain relatively unchanged, thus yielding a value for $c^{\sigma}_{u,v}$ close to 1.

Additionally, to provide a grounded comparison to the actual data, we compute the co-appearance gains $\hat{g}^{\text{low}}_{(u,v)} = \hat{p}^{\text{low}}_{u,v}/\hat{p}^{\text{low}}_{u}$ and $\hat{g}^{\text{high}}_{(u,v)} = \hat{p}^{\text{high}}_{u,v}/\hat{p}^{\text{high}}_{u}$, respectively describing how much more/less likely an urban polygon of class $u \in U$ is to be within a low or high SES tile given that it co-occurs with an urban polygon of class $v \in U$. Notice that for a pair of urban classes (u, v), neither coactivation ratios nor co-appearance gains are symmetrical. As we focus on residential areas, we constrain u to be from a very-low to very-high-density residential area (while omitting isolated residential areas because they are, by definition, less likely to co-occur with other urban classes). Coactivations and co-appearance gains for the city of Paris are shown in Fig. 6 (detailed results with explicit numerical values for all cities are provided in Supplementary Figs. 10–13). As before, we refer to an urban class u being over-coactivated in the presence of class v whenever its coactivation ratio is greater than one, that is, $c^{\sigma}_{u,v} > 1$.

Regarding the coactivations (see definition above), most dense residential areas appear over-coactivated when they neighbour clearly non-residential areas. For example, except for low-SES predictions in Marseille, very-high-density residential areas are over-coactivated when occurring with natural areas or water bodies. This also extends to agricultural areas. Interestingly, these results are observed both for high- and low-SES predictions. Very-low-density areas, on the other hand, appear over-coactivated for low-SES predictions whenever they neighbour green urban areas or leisure facilities in Paris, Lyon and Marseille. Focusing solely on residential areas, with the exception of low-SES predictions in Marseille and Nice, we observe that, whenever residential areas of different densities co-occur with one another, no over-coactivation tends to be observed. We can therefore conclude that urban classes activate differently depending on the other classes with which they co-occur, thus disproving H1. For residential areas, these differences appear when they co-occur with non-residential ones, in which case they seem to be over-coactivated, implying that our models tend to disregard pixels contained in these non-residential urban classes in favour of those within residential areas.

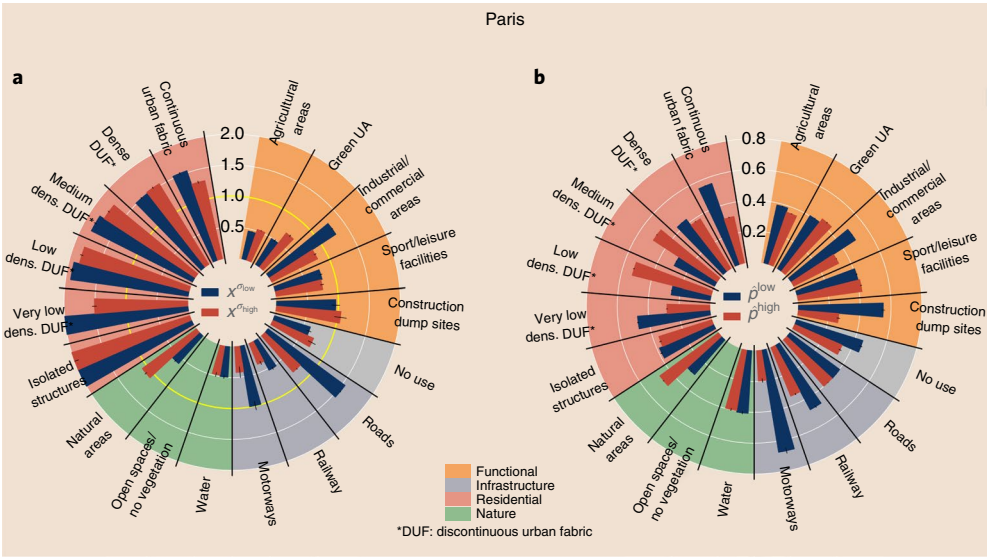


Fig. 5 | Correlations between urban topology and SES in the city of Paris. a, Mean model activation rate per urban class with error bars denoting the bootstrapped 95% confidence interval for samples predicted, respectively, as low SES (blue) or high SES (red) by the model. The yellow line on the left plot indicates the random diffusion value (H0). **b**, Estimated probability of an urban polygon belonging to the bottom or top quintile of the income distribution, with error bars denoting the bootstrapped 95% confidence interval. Sample sizes for predicted and true values per socioeconomic and urban class are provided in Supplementary Tables 2 and 3.

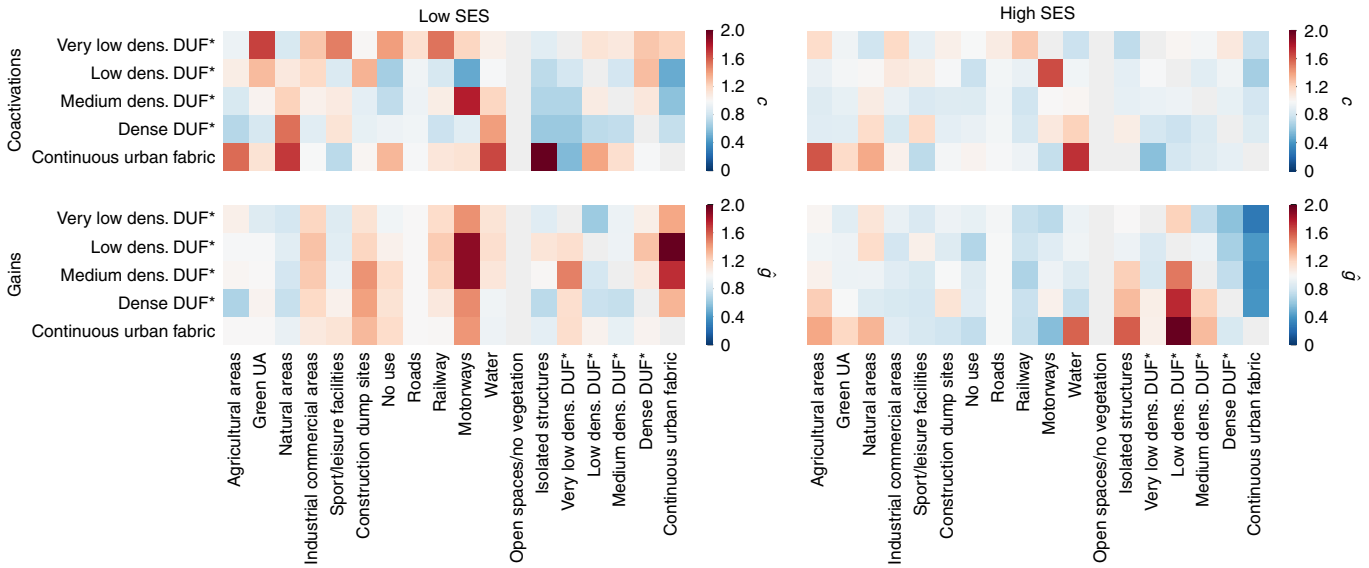


Fig. 6 | Heatmap representation of low/high SES coactivation and co-appearance gain values for the city of Paris. All residential areas of interest (rows) are listed on the left side of the map. The corresponding neighbouring areas (columns) are provided on the bottom of each map. Empty cell A_{ij} with $i \neq j$ corresponds to cases where residential area i was not seen to neighbour urban area j in at least 10 samples. Numerical values are provided in Supplementary Figs. 10–13.

This result is quite surprising, especially when we consider that the amenities that surround a given residential area are actually quite determinant in terms of the SES of the people inhabiting it. For example, when looking at the co-appearance gains in Fig. 6 and Supplementary Figs. 12 and 13, we see that low- and medium-density residential areas co-occurring next to railways and fast transit roads are more likely to be of low SES than when co-occurring with other urban classes. Similarly, very-high-density residential areas neighbouring water bodies or natural areas are more likely to be of high SES than otherwise.

Discussion

The successful deployment of urban policies aimed at curbing social problems such as income inequality, segregation and poverty requires an updated and upscaled socioeconomic description of the city. Although nationwide censuses are meant to provide such information, their prohibitive cost makes their collection rather infrequent. This, in turn, induces a lag between the actual socioeconomic changes and when they are recorded in the data. As new sources of remotely sensed imagery become available, the opportunity to frequently complement the census with socioeconomic information inferred from these digital collections has never been greater. This has been recognized even by policymakers, who rely on these novel solutions more and more despite their inherit lack of interpretability.

Our aim in this study has been to address this shortfall by (1) building a dataset from open data sources ready to be used for deep-learning-based urban solutions, (2) providing a CNN-based

framework for predicting socioeconomic status from aerial imagery for five French cities, (3) establishing a method to interpret the learned activation patterns to infer SES by mapping them back onto the original aerial images and (4) examining the activations derived from these models in terms of land cover. In doing so, we have built a model able to generate socioeconomic predictions with great performance and high spatial resolution, serving as a baseline for future studies. More importantly, we have provided a way to interpret the activation patterns of trained CNN models in terms of land use. The results show that models trained to infer SES from aerial imagery seem to rely on features contained mostly within residential areas rather than non-residential ones. Finally, we have also observed that this pattern of activations differs significantly from the statistical information one can derive by joining land cover and socioeconomic information.

Nevertheless, the interpretation of activations provided here has to be carried out with caution, as it remains unclear how the choice of architecture, training parameters or the transferability of the learned patterns in space and time change the observed correlations. More generally, the degree to which our results might inform urban theory will depend on the scale at which city-wise phenomena are considered. Therefore, although our work yields little insight into the existence of global phenomena shaping the income distribution of a given city (mainly because of the within-city income quantilization performed), it does provide evidence of how the local assembling of visual urban patterns is correlated with the income of a given area.

Indeed, our findings confirm earlier observations made across other metropolitan areas. For example, there is a considerable amount of research[27–29] showing that socioeconomic factors shape the quantity and quality of green spaces and their ability to supply services by influencing management and planning decisions. In accordance, our analysis suggests that cells containing natural areas are more likely to be wealthy, and residential areas neighbouring natural areas are less likely to be poor. A similar claim could be made regarding the link between motorways and lower SES. In doing so, our work paves the way to develop more comprehensive urban theories where both urban perception and topology are used to understand the distribution of income within cities.

Nevertheless, relying on land-use maps to interpret our models is not without its challenges. Our framework relies on three layers of data, which might not be easily obtainable in developing countries, thus making other approaches more suitable for interpretable SES prediction[13,15,15,16] in these settings. However, our work is particularly well placed to study urban poverty across a wide range of modern metropolises that are covered by either the present or similar datasets. Moreover, as higher-resolution satellite coverage becomes more available and computer vision advances for land cover segmentation become better performing, these datasets might become more prevalent in the future, further enhancing the applicability of this work. These caveats indicate the research avenues that need to be followed to explore different models, resolutions and interpretability techniques to reach the end goal of deploying these models with a more complete understanding of their inner mechanisms.

Our research direction is nevertheless one of many, as new works are attempting to increase the interpretability of such systems by detecting individual objects in satellite images[15] or relying on a different type of data altogether that lends itself better to be interpreted[16], as well as increasing their transferability[13,30].

To contribute to solving these challenges, we have made all our code and data pointers openly available to the community. We hope this proof of concept motivates further work in this field, where the identification of global patterns of poverty could have far-reaching impact by enabling the swift allocation of needed resources to the most vulnerable.

Broader impact

Decision-makers are increasingly turning to technological solutions to inform or even devise their policies. The surge in demand for artificial intelligence (AI)-powered tools can be met as long as the models they rely on provide some accountable insights into their behaviour. Our paper addresses some of these issues by building a joint dataset and deep-learning framework where SES predictions from aerial images can be interpreted in terms of urban planning. In doing so, it provides firm ground on which to study how a city's local structure can be informative about the distribution of wealth in it. Furthermore, as new deep-learning architectures continue to emerge, it paves the way to build more advanced systems for SES inference, which can then be applied, analysed and interpreted using a common framework. Finally, similarly to what is currently taking place in the medical imaging community, we believe our work can enable the design of more sophisticated urban solutions that are poised to integrate expert knowledge from urban planners as useful prior knowledge to feed to deep-learning models.

There are, however, many missing key elements that need to be resolved before we can take full advantage of these systems. Interpretation of deep-learning models is currently an area of active research where standard benchmarks for interpretability are yet to be found. Despite the abundance of available methods, it is still hard to judge what are the salient features in an image that would make a given model more interpretable. Our contribution takes a step in this direction by proposing a simple framework for the evaluation of interpretability in SES inference.

Moreover, biases present in training data may lead models to generate prejudiced content. This is concerning, because model biases could harm people in lower socioeconomic classes in different ways by entrenching existing stereotypes and producing demeaning portrayals, for example. It is therefore inadvisable to think that such models can be used in their present state as tools to inform the allocation of resources to economically deprived communities.

Methods

Our work was set up to address (1) the inference of the SES associated with a given location from its aerial image and (2) the derivation of saliency maps for these images to find correlations between the activation patterns of the models and the underlying urban structure.

Model description for SES inference. Our model used the original EfficientNetB0 (EB0) architecture, consisting of a series of mobile inverted bottleneck blocks (MBConv) augmented with squeeze-and-excitation optimization. In addition, we added a max pooling layer to downsample the feature maps, followed by global average pooling and dense layers (Supplementary Fig. 1). To yield an estimated probability of income quantiles for each input, we followed the approach used by Suel et al.[12], and connected the dense layer to a binomial layer taking a single p-value, interpreted as a probability with which Bernoulli trials are performed, as input and outputting the probabilities for each SES class $Y = [\hat{y}_1, ..., \hat{y}_{n_{\mathrm{SES}}}]$ with

$$\hat{y}_k = \binom{n_{\mathrm{SES}}}{k-1} p^{k-1}(1-p)^{n_{\mathrm{SES}}-(k-1)}, \quad k \in [1, ..., n_{\mathrm{SES}}]$$

In doing so, the model becomes sensitive to the ordinal relationship existing between quantiles. Training is then conducted with a classic categorical cross-entropy loss.

It is worth mentioning that the original EB0 model was trained on object-centric $224 \times 224 \times 3$ images and its initial weights were therefore not conceived to handle $800 \times 800 \times 3$ aerial images by default. However, because of the model's fully convolutional nature, these weights could still be used to kickstart the training process. The model was thus trained by transfer learning from ImageNet and without freezing any layer. This of course assumes that the optimization procedure will more readily converge by adapting the weights from the optimal ones learned at lower resolutions than by actually learning them from scratch. This assumption is reasonable, because the first layers of the network will already be sensitive to low-level features.

In all our experiments, we augmented the data through random horizontal and vertical flipping of the input images. Images were scaled to $800 \times 800 \times 3$ to avoid excessively decreasing the original image resolution while keeping the use of our computational resources economic. An Adam variant of stochastic gradient descent was used to learn the model weights. Initially, we started with a learning

rate of 8×10^{-5} and decreased it by 90% every three epochs until the validation loss showed no improvement for more than 10 epochs. All networks were trained on either NVIDIA Tesla K20X or V100 graphics processing units for at most 30 epochs, with 2,500 samples in each epoch. Furthermore, within each city, we used fivefold cross-validation. In each fold, 80% of data were used for training the network and the remaining 20% were withheld. The training set was further randomly subdivided with a 75–25% split for inner-fold training and validation. Every reported performance metric was then averaged over all folds.

Finally, while recent work has shown that the prediction on satellite images can be further improved by combining the original images with land-use masks[31], we decided against this design for the sake of the interpretability of the model: by inputting land-use information into the model, the backtracking of the network's activations performed by Grad-CAM would no longer be agnostic with respect to the urban topology of the image. This would then make salient features returned by this method no longer solely dependent on the image itself, but on the urban features as well, thus biasing the interpretation of the model's activations overall (see next subsection).

Guided Grad-CAM. Guided Grad-CAM follows the CNN's gradient flow from individual output classes back onto the original image tile to establish an activation map, highlighting the input features most relevant to each CNN prediction. To generate such mappings, we first computed saliency maps via guided backpropagation, producing a pixel-space gradient map of predicted class scores with respect to the pixel intensities of the input image. This gradient visualization, though fine-grained, is known not to be class-discriminative. It can, however, be combined with Grad-CAM via pointwise multiplication to generate guided Grad-CAM (see Supplementary Information for more details). Grad-CAM itself results from the dot product of the feature map of the last convolutional layer and the partial derivatives of predicted class scores with respect to the neurons in the last convolutional layer. Incidentally, it is this last task that guided the design of our original socioeconomic inference model: because of its small number of parameters, EB0 allows for a fast computation of activation maps while still performing on par with more classical models (Supplementary Table 1). Furthermore, because of its fully convolutional nature, aerial images can be downsampled deeper within the network (at the global average pooling layer), rather than prior to being fed to the model, enabling us to obtain 800×800-pixel activation maps. This in turn grants us the possibility of taking full advantage of the high resolution of the aerial images (originally $1{,}000\times1{,}000$ pixels) and land-use maps to obtain a more fine-grained view of land-use-to-income correlations. All mappings were done in a cluster of 96 CPUs.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

In this paper we built on the combination of three publicly available datasets. One was issued by the National Geographical Information Institute (IGN) and contains aerial images about French municipalities[20]. The second was provided by the French National Institute of Statistics and Economic Studies (INSEE) in 2019[21] and provides a high-resolution socioeconomic map. The third was shared by the European Environment Agency through the 2012 European Union Urban Atlas project about EU28 and EFTA countries[22]. All datasets are public and openly accessible online at https://geoservices.ign.fr/documentation/diffusion/telechargement-donnees-libres.html#ortho-hr-sous-licence-ouverte, https://www.insee.fr/fr/statistiques/4176290?sommaire=4176305 and https://land.copernicus.eu/local/urban-atlas/urban-atlas-2012. The individual files that were downloaded from the aerial imagery dataset are provided in GitHub. Figures depicting raw data are shown in Figs. 1, 3 and 4 and Supplementary Fig. 1. All figures and tables are provided as source data at https://doi.org/10.6084/m9.figshare.12595067.v2.

Code availability

Code developed for the research has been made freely available for non-commercial use under an MIT Licence and shared through an open repository[32] at https://doi.org/10.5281/zenodo.3906063. For any further questions please contact the corresponding authors.

Received: 28 February 2020; Accepted: 25 September 2020;
Published online: 26 October 2020

References

1. *World Urbanization Prospects*: *The 2018 Revision* (United Nations, 2019).
2. Glaeser, E. L. & Joshi-Ghani, A. *The Urban Imperative*: *Towards Competitive Cities* (Oxford Univ. Press, 2015).
3. Gourevitch, M. N., Athens, J. K., Levine, S. E., Kleiman, N. & Thorpe, L. E. City-level measures of health, health determinants, and equity to foster population health improvement: the city health dashboard. *Am. J. Public Health* **109**, 585–592 (2019).
4. *Revenus, Pauvreté et Niveau de Vie en 2014* (INSEE, 2014).
5. Seto, K. C. & Kaufmann, R. K. Modeling the drivers of urban land use change in the Pearl River Delta, China: integrating remote sensing with socioeconomic data. *Land Econ.* **79**, 106–121 (2003).
6. Stead, D. Relationships between land use, socioeconomic factors and travel patterns in Britain. *Environ. Planning B Planning Des.* **28**, 499–528 (2001).
7. Mirmoghtadaee, M. The relationship between land use, socio-economic characteristics of inhabitants and travel demand in new towns—a case study of Hashtgerd New Town (Iran). *Int. J. Urban Sustain. Dev.* **4**, 39–62 (2012).
8. Kinzig, A. P., Warren, P., Martin, C., Hope, D. & Katti, M. The effects of human socioeconomic status and cultural characteristics on urban patterns of biodiversity. *Ecol. Soc.* **10**, 23 (2005).
9. Blumenstock, J. E. Fighting poverty with data. *Science* **353**, 753–754 (2016).
10. Llorente, A., Garcia-Herranz, M., Cebrian, M. & Moro, E. Social media fingerprints of unemployment. *PLoS ONE* **10**, 1–13 (2015).
11. Dong, L., Ratti, C. & Zheng, S. Predicting neighborhoods' socioeconomic attributes using restaurant data. *Proc. Natl Acad. Sci. USA* (2019); https://doi.org/10.1073/pnas.1903064116
12. Suel, E., Polak, J. W., Bennett, J. E. & Ezzati, M. Measuring social, environmental and health inequalities using deep learning and street imagery. *Sci. Rep.* **9**, 6229 (2019).
13. Jean, C. et al. Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
14. Timnit, G. et al. Using deep learning and Google street view to estimate the demographic makeup of neighborhoods across the United States. *Proc. Natl Acad. Sci. USA* **114**, 13108–13113 (2017).
15. Ayush, K., Uzkent, B., Burke, M., Lobell, D. & Ermon, S. Generating interpretable poverty maps using object detection in satellite images. In *Proc. Twenty-Ninth Int. Joint Conf. Artificial Intelligence* 4410–4416 (International Joint Conferences on Artificial Intelligence Organization, 2020).
16. Sheehan, E. et al. Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (ACM, 2019); https://doi.org/10.1145/3292500.3330784
17. Selvaraju, R. R. et al. Grad-cam: visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (*ICCV*) 618–626 (Springer, 2017); https://doi.org/10.1109/ICCV.2017.74
18. Iizuka, T., Fukasawa, M. & Kameyama, M. Deep-learning-based imaging-classification identified cingulate island sign in dementia with Lewy bodies. *Sci. Rep.* **9**, 8944 (2019).
19. Tang, Z. et al. Interpretable classification of alzheimer's disease pathologies with a convolutional neural network pipeline. *Nat. Commun.* **10**, 2173 (2019).
20. *Données ORTHO HR* (IGN, accessed 1 March 2020); https://geoservices.ign.fr/documentation/diffusion/telechargement-donnees-libres.htm l#ortho-hr-sous-licence-ouverte
21. *Données carroyées* (INSEE, accessed 1 March 2020); https://www.insee.fr/fr/statistiques/4176290?sommaire=4176305
22. *Urban Atlas 2012* (Copernicus Land Monitoring Service, accessed 1 March 2020); https://land.copernicus.eu/local/urban-atlas/urban-atlas-2012
23. Tan, M. & Le, Q. EfficientNet: rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* Vol. 97, 6105–6114 (PMLR, 2019).
24. Hosny, A. et al. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. *PLoS Med.* **15**, 1–25 (2018).
25. Miao, Z. et al. Insights and approaches using deep learning to classify wildlife. *Sci. Rep.* **9**, 8137 (2019).
26. Toda, Y. & Okura, F. How convolutional neural networks diagnose plant disease. *Plant Phenomics* (2019); https://doi.org/10.1155/2019/9237136
27. Shanahan, D. F., Lin, B., Gaston, K., Bush, R. & Fuller, R. Socio-economic inequalities in access to nature on public and private lands: a case study from Brisbane, Australia. *Landscape Urban Planning* **130**, 14–23 (2014).
28. Wilkerson, M. L. et al. The role of socio-economic factors in planning and managing urban ecosystem services. *Ecosyst. Services* **31**, 102–110 (2018).
29. You, H. Characterizing the inequalities in urban public green space provision in Shenzhen, China. *Habitat Int.* **56**, 176–180 (2016).
30. Head, A., Manguin, M., Tran, N. & Blumenstock, J. E. Can human development be measured with satellite imagery? In *Proc. Ninth International Conference on Information and Communication Technologies and Development*, ICTD '17 (Association for Computing Machinery, 2017); https://doi.org/10.1145/3136560.3136576
31. Audebert, N., Le Saux, B. & Lefevre, S. Joint learning from Earth observation and openstreetmap data to get faster better semantic maps. In *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) *Workshops* 1552–1560 (IEEE, 2017); https://doi.org/10.1109/CVPRW.2017.199

32. Abitbol, J. L. Coding repository; https://github.com/jaklevab/SESEfficientCAM/.

Acknowledgements
This work was partially funded by the SoSweet ANR project (ANR-15-CE38-0011), the MOTIf Stic-AmSud project (18-STIC-07) and the ACADEMICS project financed by IDEX LYON. M.K. was supported by the DataRedux (ANR-19-CE46-0008) ANR and the SoBigData++ (871042) H2020 projects.

Author contributions
J.L.A. and M.K. designed the research. J.L.A. built the combined dataset and implemented analysis of the results. J.L.A. and M.K. wrote the final manuscript.

Competing interests
The authors declare no competing interests.

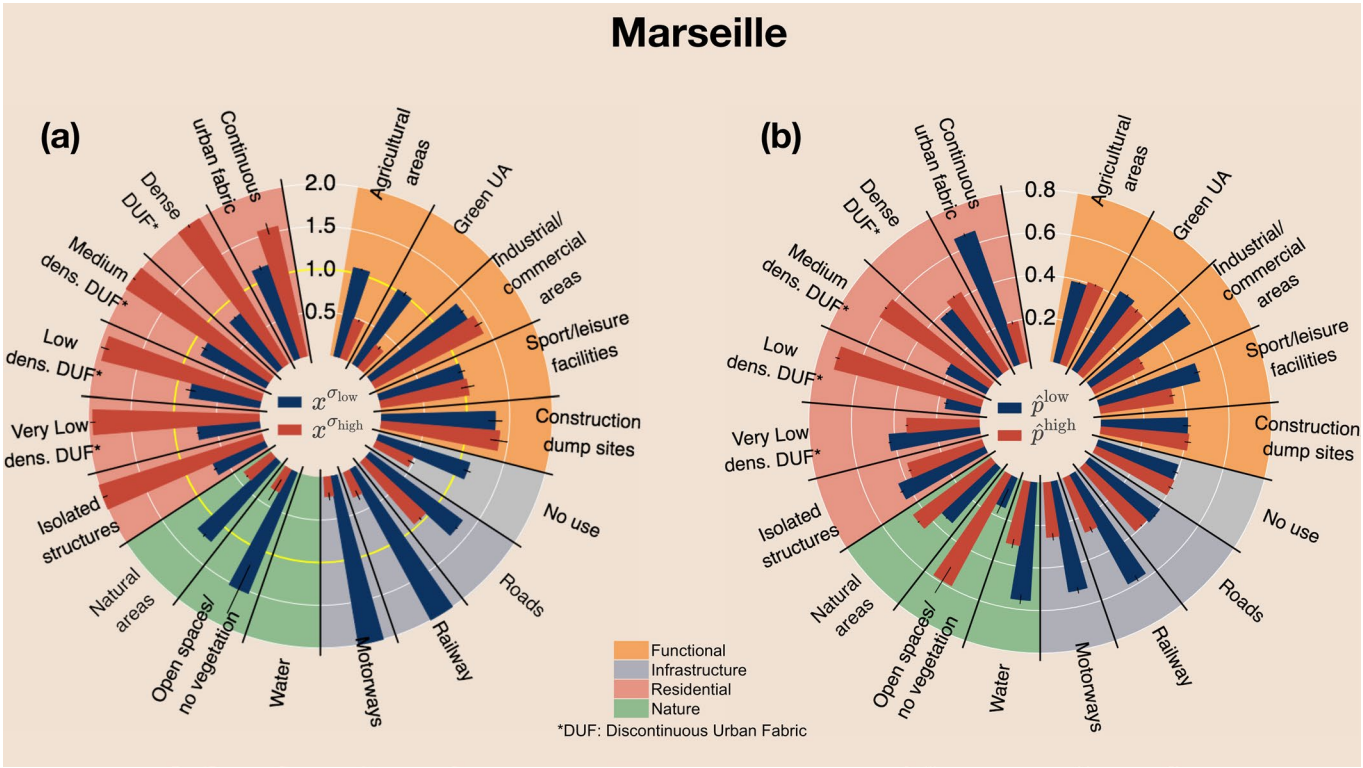
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s42256-020-00243-5.

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-020-00243-5.

Correspondence and requests for materials should be addressed to J.L.A. or M.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Correlations between urban topology and socioeconomic status in the city of Marseille. a, Mean model activation rate per urban class with error bars denoting the 95% confidence interval for samples predicted as respectively as low SES (blue) or high SES (red) by the model. The yellow line on the left plot indicates the random diffusion value (H0). b, Estimated probability of an urban polygon belonging to the bottom or top quintile of the income distribution with error bars denoting the 95% confidence interval. Sample sizes are provided in Supplementary Tables 2 and 3.

nature research

Corresponding author(s): Jacobo Levy Abitbol and Márton Karsai

Last updated by author(s): Mar 12, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Aerial Images of the five metropolitan areas at hand were obtained from the National Geographical Information Institute (IGN) HR Orthophotography collection. High ResolutionSocioeconomic maps were provided by the French National Institute of Statistics and Economic Studies (INSEE) in the 200m x 200m gridded dataset collection. Urban Atlases were obtained for each metropolitan area from the European Environment Agency 2012 European Union Urban Atlas project for all EU28 and EFTA countries. All datasets are public and openly accessible online.
Data analysis	Code developed for the actual research is shared through an open repository https://github.com/jaklevab/SESEfficientCAM along with some sample data. EfficientNet models were built upon based on the following implementation:https://github.com/qubvel/efficientnet GradCAM and Guided GradCAM are provided within the aforementioned repository. Stitching of aerial tiles, extraction of individual images corresponding to census cells and aggregation of gradcam values based on urban polygons were performed using the shapely, rasterio and rasterstats python libraries. All further implementations relied on standard python libraries (pandas, numpy, sklearn, scipy and keras)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All datasets upon which this study was conducted are public and openly available online (only for non- commercial purposes in the case of the aerial image

collection). Figures 1, 3, 4, and 7 in the main paper illustrate some raw data samples.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We aimed to build a high resolution model for socioeconomic status inference on aerial imagery and dissect the model's activations in terms of the underlying land cover classes.
Research sample	Our sample contains the five French metropolitan areas containing the most population and census cells (synonymous for land cover area).
Sampling strategy	All cells contained within the boundaries set by the EUEA for metropolitan areas were included in our dataset.
Data collection	Data collection was performed as indicated in the previous section.
Timing and spatial scale	The spatial scale varies from city to city ranging from 474 (for Nice) to 2720 square kilometers (for Paris). The spatial scale of aerial images were 20cm/pixel, the spatial resolution of the socioeconomic map was 200m x 200m. The timing corresponds to a time-window from 2013 to 2016 in which all of the aforementioned datasets were collected.
Data exclusions	When considering the variations of activation patterns within different urban classes, port areas were excluded as they didn't provide enough data points on which to aggregate the activation values. Other urban classes contained in the European Environment Agency dataset were merged according to the procedure laid out in Albert et al. (2017), see reference in the manuscript.
Reproducibility	Reported results were consistently replicated across multiple runs with all replicates generating similar results.
Randomization	We do not use randomization to proof or disproof observations, but only to define reference models for the comparison of our observations. Randomization under a case control scenario is not generally used in this field.
Blinding	Investigators were not blinded. Blinding during collection was not needed as conditions were well controlled. Blinding during analysis was not necessary as our models converged to similar optima among different training runs. Blinding is not typically used in the field.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging