Predicting City Poverty Using Satellite Imagery

Simone Piaggesi∗† Laetitia Gauvin∗ Michele Tizzoni∗ Natalia Adler‡ Stefaan Verhulst§
Andrew Young§ Rihannan Price¶ Leo Ferres∥ Ciro Cattuto∗ André Panisson∗

Abstract

Reliable data about socio-economic conditions of individuals, such as health indexes, consumption expenditures and wealth assets, remain scarce for most countries. Traditional methods to collect such data include on-site surveys that can be expensive and labour intensive. On the other hand, remote sensing data, such as high-resolution satellite imagery, are becoming largely available. To circumvent the lack of socio-economic data at high granularity, computer vision has already been applied successfully to raw satellite imagery sampled from resource poor countries.

In this work we apply a similar approach to the metropolitan areas of five different cities in North and South America, starting from pre-trained convolutional models used for poverty mapping in developing regions. Applying a transfer-learning process we estimate household income from visual satellite features. The urban environment we consider is characterized by different features with respect to the resource-poor training environment, such as the high heterogeneity in population density. By leveraging both official and crowd-sourced data at city scale, we show the feasibility of estimating the socio-economic conditions of different neighborhoods from satellite data.

1. Introduction

For years, estimating economic growth and development to assess human well-being has been a central issue for international research and policy [3, 11]. Accurate data about human development primarily come from surveys and censuses. Collecting socioeconomic information can be hardworking, expensive and might suffer from reporting errors.

In the last years, the huge amount of remote sensing data, combined with recent developments in machine learning techniques, led to methods for estimating socio-economic indicators from geospatial data, such as nightlights [1, 17] and satellite imagery [26]. In particular, such methods have been used to overcome the limitations or lack of development data and to estimate poverty at large scale in developing countries [22, 9, 16].

In this work we apply a machine learning approach similar to the one developed by Jean et al. [16] for predicting economic outcomes in 5 African countries from satellite imagery, to the smaller scale of an urban environment. In particular, we explore the feasibility of predicting household income at various municipality levels in a city. To this aim, we examined the Metropolitan Area of Santiago in Chile and other five big cities in the USA: Los Angeles, Philadelphia, Boston, Chicago, and Houston. Our work tackles three main research questions:

1. Is it possible to extend machine learning methods, previously applied to resource poor settings, to estimate poverty levels in a city of a developed country?

2. Given different aggregation levels in a city – usually corresponding to different administrative subdivisions – can a model trained on a lower spatial resolution yield information about a more granular aggregation level?

3. What is the out-of-sample predictive power of such a model, when tested on a new city?

Our main findings related to the above questions are:

• Starting from pre-trained deep computer vision models we show the feasibility of predicting household income in a city through a regression task, with best results on settings with only urban areas. We show that in this framework there is no need to fine tune existing models or to leverage on proxy variables (such as night-time lights data) to achieve good performances.

• Just considering municipalities in which the regressor is better predictive about the target, we can also improve the estimation in more fine-grained levels of aggregation.
2. Related Literature

In the last decade the development of new machine learning techniques, such as Deep Learning [19] and Convolutional Neural Networks [18], has allowed to extraordinarily improve many Computer Vision applications [25, 7, 12]. Recently, deep learning has been used in combination with remote sensing data for various tasks, such as Scene Classification [21, 20, 4], Urban Planning [2] and Crop Yield Prediction [27]. A recent application concerns the estimation of socioeconomic indicators, such as assets, consumption and wealth indexes, from satellite imagery. In [16] the authors proposed a transfer learning process, where nightlights intensities are used as an intermediate proxy [1] to map poverty in five African countries. Other works have used similar transfer learning approaches to estimate other variables [14], with different proxies [23] or deep models [6]. Also, other studies [24, 8, 15] have trained a deep neural network to predict poverty from satellite images without proxies, or used other types of remote sensing data for the same task [10]. All the latter models were trained and tested in resource poor settings (Africa, India, Bangladesh, and Sri Lanka) mainly considering their rural areas. To our knowledge no previous work has shown the application of similar techniques to the urban areas of a developed country.

3. Data

3.1. Economic Variables

In this work, we focus on the household income as the main economic indicator to be predicted. Data about this indicator comes from different sources, depending on the city under study, and the same indicator has to be available at different levels of granularity for validation. The following cities were selected, due to data availability:

Santiago, Chile The household income is obtained from the EOD (Encuesta Origen Destino de Viajes), a mobility survey realized from July 2012 to November 2013 by assignment of the Chilean Ministry of Transport and Telecommunications1. The survey refers to a random sample of 18,264 households coming from the Santiago Metropolitan Area, for a total of 60,054 people involved. Household income information is averaged at the municipality-level of comunas, but also at a more fine-grained city subdivision, called zonas.

USA cities Households income data are available from Census Reporter2, a website through which the United States Census Bureau provides socio-economic and demographic data on the population of the United States. The geographic levels we considered include the ZIP codes and the more fine grained census tracts.

In the following, we will refer to the different levels of aggregation (comunas, zonas, ZIP codes or census tracts) with the term clusters.

3.2. Satellite Imagery

Santiago, Chile Satellite images are downloaded as a mosaic of 34 big tiles from the DigitalGlobe web platform3, to entirely cover the area of the city. Tiles are a mixture of pansharpened and natural color RGB images, taken during daytime between September 2017 and February 2018, with 50 cm resolution and maximum cloud coverage of 3%. We generated a grid of not overlapping images from the mosaic of tiles covering the city area. The grid has a 1km step and each image is cropped from the original mosaic and resized to 400 x 400 pixels.

USA cities Satellite images are downloaded using the Google Static Maps API, each of them with 400 x 400 pixels at zoom level 16, resulting in images with a resolution of 2.5 m/pixel (1km² area per image) not overlapping each other.

Table 1 provides a summary of the geographic elements and the different spatial resolutions considered in our study. A description of the resolution scales that we are covering is reported in Fig. 1, with the distribution of the surfaces areas of the geospatial elements for each aggregation level.

3.3. Urban Areas Boundaries

Shapes and boundaries of urban areas of each city are obtained from official4 sources and then merged with OpenStreetMap5 (OSM) crowd-sourced data to have a more comprehensive description of the urban landscape. From OSM map elements, we selected urban spaces joining those with the tag key landuse with possible values: residential, commercial, retail, recreation_ground, construction, college, university, public, allotments, churchyard, depot.

4. Methods

4.1. Convolutional Features Extraction

To estimate socio-economic indicators from satellite imagery, we leverage the transfer learning approach intro-
Table 1: Number of satellite images and geospatial clusters for each city.

<table>
<thead>
<tr>
<th>Geographic Level</th>
<th>Santiago</th>
<th>Geographic Level</th>
<th>Los Angeles</th>
<th>Philadelphia</th>
<th>Boston</th>
<th>Chicago</th>
<th>Houston</th>
</tr>
</thead>
<tbody>
<tr>
<td>Images</td>
<td>12444</td>
<td>Images</td>
<td>16774</td>
<td>19528</td>
<td>14766</td>
<td>31042</td>
<td>37353</td>
</tr>
<tr>
<td>Zonas</td>
<td>588</td>
<td>Census Tracts</td>
<td>1772</td>
<td>1334</td>
<td>907</td>
<td>1829</td>
<td>1061</td>
</tr>
<tr>
<td>Comunas</td>
<td>52</td>
<td>ZIP Codes</td>
<td>290</td>
<td>359</td>
<td>282</td>
<td>403</td>
<td>237</td>
</tr>
</tbody>
</table>

Figure 1: Distribution of different cluster-level areas for the considered cities.

Produced and elaborated by [26, 16]. Such approach consists in using a Convolutional Neural Network (CNN), trained on a nighttime light intensity prediction task, as a feature extractor that maps from each input image to a vector representation, which incorporates nightlights features. Such features are proved to be a good proxy for economic development [17, 1], and can be used at the bottom of a linear regression model to predict wealth-related indicators.

To perform the forward pass from raw satellite images to visual features we have considered three CNN models:


- **VGGF+nightlights.** The fully convolutional variant of the previous one, fine-tuned with nightlights intensity labels on some African countries (Uganda, Tanzania, Nigeria, Malawi).

In essence, we compared a model fine-tuned on the nightlights prediction task with other two models (VGGF and ResNet50) just pre-trained on the ImageNet dataset, with the aim of quantifying the goodness of fit when using general features not related with night-time lights.

For each tile, we extract vector representations from the top layer of the model, before the softmax classifier. The size of our images is the same respect to the input of the VGGF fine-tuned model, while the pre-trained neural networks take as input 224x224 pixels images. In this case, we averaged the feature vectors extracted from the four 224x224 overlapping quadrants of each image to get vector representations. In addition, each image is rotated by multiples of 90 degrees and flipped horizontally/vertically before the mapping, to obtain an augmented dataset.

Figure 2: Diagram with the collected datasets, feature extraction and data preparation processes for the experiments. The feature vector \( x_t \) is obtained from a CNN model using a satellite image as input. \( y_{ti} \) and \( y_{zi} \) are the socio-economic variables extracted from two different levels of census area boundaries, associated to the original satellite image.

The diagram on Fig. 2 shows the collected datasets, feature extraction and data preparation processes. For each city, satellite images are divided in a grid and each tile is associated to socio-economic variables at different levels of census area boundaries. A vector representation for each tile is extracted with different CNN models, as explained.
above. For each experiment in Section 5, a different part of this dataset is used as training and validation sets.

4.2. Regression for Spatial Prediction

As observed in previous applied computer vision works, visual features can be used to estimate socio-economic indicators. Here we test the prediction of the household income variable on multiple aggregation levels in different cities.

To explore how features inferred from the urban environment influence the final estimations, we can train and validate our models on different subsets of images, belonging or not to urban areas. In this section every image is assigned to an urban area if its center falls inside the shape of the area, but we have observed that assigning to urban regions also images with marginal overlaps with the latter (less restrictive statement respect the one above) does not change considerably final the results. Urban boundaries are obtained following the methodology explained in Sec. 3.

We perform each experiment with a Ridge Regression model, using image vector representations as predictor variables and income values as target. Hyperparameters are estimated with a 10-fold cross-validation, using as metric the determination coefficient $R^2$ and evaluating it on different validation sets (specified in each experiment).

We distinguish two different cases of prediction:

**Image-level Estimation** We use features of image-level embeddings as set of predictors and assign to this the target value up-sampled to an individual resolution level (comunas, zonas, ZIP code or census tract). The absence of superimposition among images prevents the information leaking between training and test sets in the learning process. Because of the fine-grained property of the set of images, in this case it is possible to tune the model on points that belong or not to urban areas. Then, we can investigate how much the presence of urban areas may influence the final performance of the model, but also see if a global training improves the estimation only in urban spaces.

**Cluster-level Estimation** For each cluster we use as predictor variables the average of image-level embeddings computed on the images which are part of the cluster itself. This is the same approach used in previous published works. As done by [16], we neglect spatial clusters with less than 10 images for the regression task.

5. Experiments

5.1. Income Prediction within Cities

The first experiment concerns the income estimation at the level of a single city, to understand two significant aspects: which convolutional model yields more informative features from images, and what is the importance of the level of urbanization in terms of regression performance. Table 2 shows the results for the municipality of Santiago, for brevity we omit outcomes for all the other cities because they are equally demonstrative. We can observe that globally the pre-trained ResNet50 model performs better at every considered resolutions, besides the VGGF model fine-tuned on nightlights has low performances in any task (meaning that this model is not reliable if not applied to the original data space where it is fine-tuned). For this evidence, we perform all the other experiments presented with satellite features extracted with the ResNet50 model.

Table 3 illustrates the analysis of urban areas, using only the ResNet50 features (having observed that it is the best model among all those considered). Here, features are image-level and we compute the regression score in two validation sets: a first one with all the images and a second one including only those related to urban spaces. We can observe that in most cases the regressor can better explain the variance of the target in urban neighborhoods. Fig. 3 displays the spatial heatmaps comparing the true income distribution with the predicted one in the city of Santiago.

![Figure 3: Spatial representations for the image-level prediction task performed in Santiago. Visual features are extracted with the ResNet50 model, the target variable is the zonas-level household income. Left: Distribution of real values. Right: Distribution of predicted values.](image)

5.2. Inference of Higher Resolution Estimates

In Table 3 we have seen that we can estimate the distribution of the household income, taking individual image features as input, at a more fine-grained resolution with respect to the target itself (each image is $1km^2$). In this section we investigate more in depth this aspect, asking if with a training step on the more coarse-grained level the regression model can also assess the distribution of the target at a more fine-grained level.

To this aim, we consider two sets of aggregation levels $T$ and $Z$ of the same geographical area (suppose $T$ to be more
Table 2: Image-level and cluster-level (zonas and comunas) cross validation performance, measured with $R^2$ scores, of the household income regression task in Santiago. Here we compare features coming from different convolutional models (ResNet50, VGGF, VGGF+nightlights). The ResNet50 features outperform the others in almost all tasks. Training and/or validation sets are coupled with different neighborhoods of the city (urban areas or not).

<table>
<thead>
<tr>
<th>Features Level</th>
<th>Target Level</th>
<th>Train.-Val.</th>
<th>ResNet50</th>
<th>VGGF</th>
<th>VGGF+nightlights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zonas</td>
<td>Zonas</td>
<td>All Areas</td>
<td>0.477</td>
<td>0.484</td>
<td>0.356</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Urban Areas</td>
<td>0.591</td>
<td>0.523</td>
<td>0.433</td>
</tr>
<tr>
<td>Comunas</td>
<td>Comunas</td>
<td>All Areas</td>
<td>0.643</td>
<td>0.598</td>
<td>0.553</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Urban Areas</td>
<td>0.737</td>
<td>0.711</td>
<td>0.623</td>
</tr>
<tr>
<td>Images</td>
<td>Zonas</td>
<td>All - All</td>
<td>0.454</td>
<td>0.408</td>
<td>0.258</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All - Urban</td>
<td>0.520</td>
<td>0.506</td>
<td>0.314</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Urban - Urban</td>
<td>0.584</td>
<td>0.542</td>
<td>0.358</td>
</tr>
<tr>
<td>Images</td>
<td>Comunas</td>
<td>All - All</td>
<td>0.667</td>
<td>0.613</td>
<td>0.342</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All - Urban</td>
<td>0.691</td>
<td>0.625</td>
<td>0.429</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Urban - Urban</td>
<td>0.772</td>
<td>0.713</td>
<td>0.503</td>
</tr>
</tbody>
</table>

Table 3: Image-level cross validation performance, measured with $R^2$ scores, with different environments in the validation set. The training set includes image-level features extracted with the ResNet50 model, from both urban and rural areas. We observe that the validation score is higher on urban areas, regardless of the granularity level of the target.

<table>
<thead>
<tr>
<th>City</th>
<th>Target Level</th>
<th>Validation set</th>
<th>All</th>
<th>Urban</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santiago</td>
<td>Zonas</td>
<td>0.454</td>
<td>0.520</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Comunas</td>
<td>0.667</td>
<td>0.691</td>
<td></td>
</tr>
<tr>
<td>Los Angeles</td>
<td>Census Tracts ZIP Codes</td>
<td>0.657</td>
<td>0.569</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.569</td>
<td>0.458</td>
<td></td>
</tr>
<tr>
<td>Philadelphia</td>
<td>Census Tracts ZIP Codes</td>
<td>0.360</td>
<td>0.460</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.358</td>
<td>0.443</td>
<td></td>
</tr>
<tr>
<td>Boston</td>
<td>Census Tracts ZIP Codes</td>
<td>0.384</td>
<td>0.374</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.367</td>
<td>0.399</td>
<td></td>
</tr>
<tr>
<td>Chicago</td>
<td>Census Tracts ZIP Codes</td>
<td>0.301</td>
<td>0.361</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.309</td>
<td>0.382</td>
<td></td>
</tr>
<tr>
<td>Houston</td>
<td>Census Tracts ZIP Codes</td>
<td>0.250</td>
<td>0.327</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.266</td>
<td>0.340</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Cross-validation performance for the cluster-level income prediction task, when the the Ridge Regression is trained to predict ZIP codes-level (or comunas) target values, but validated on census tracts-level (or zonas) set of target values. Scores are compared with a baseline computed assigning to each census tract (or zona) the target of the corresponding ZIP code (or comuna).

<table>
<thead>
<tr>
<th>City</th>
<th>score</th>
<th>baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santiago</td>
<td>0.411</td>
<td>0.483</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>0.531</td>
<td>0.618</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>0.631</td>
<td>0.684</td>
</tr>
<tr>
<td>Boston</td>
<td>0.523</td>
<td>0.626</td>
</tr>
<tr>
<td>Chicago</td>
<td>0.447</td>
<td>0.706</td>
</tr>
<tr>
<td>Houston</td>
<td>0.506</td>
<td>0.614</td>
</tr>
</tbody>
</table>

Our purpose consists in training the regression model with cluster-level features on the lower resolution level (i.e. the set $Z$), but maximizing the validation score on the higher one (i.e. the set $T$). From Table 4 we can argue that this procedure does not improve the baseline.

We can go beyond, validating only on areas in which the model predictions are better. To do so, firstly we sorted the lower level clusters according to a local regression loss (the squared error between the actual target and the predicted one), and secondly we introduced an additional hyperparameter $q$ (optimized during the cross-validation) which tunes the fraction of clusters taken into account for the validation.

In this way, we can also estimate the number of ele-
City | \( q \) | \( \text{score}_@q \) | \( \text{baseline}_@q \)
--- | --- | --- | ---
Santiago | 0.35 (39\%) | 0.484 | 0.481
Los Angeles | 0.1 (0\%) | 0.548 | 0.616
Philadelphia | 0.25 (28\%) | 0.628 | 0.617
Boston | 0.1 (0\%) | 0.645 | 0.761
Chicago | 0.2 (0\%) | 0.454 | 0.598
Houston | 0.4 (39\%) | 0.578 | 0.542

Table 5: Cross-validation performance for the cluster-level income prediction task, when the the Ridge Regression is trained as in Table 4, but validated on census tracts (or zonas) belonging to the fraction \( q \) of best predicted ZIP codes (or comunas). The choice of the hyperparameter \( q \) is optimized during the cross-validation, between values from 0.1 and 0.5 with step 0.05. In parentheses are shown corresponding percentages of higher-level clusters (census tracts or zonas) for which the prediction is improved respect to the baseline by the regression model. Baseline values are different respect to Table 4 because are evaluated on a different subset of clusters.

<table>
<thead>
<tr>
<th>City</th>
<th>Target Level</th>
<th>Our Model</th>
<th>Null Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>Urban</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>Census Tracts</td>
<td>0.435</td>
<td>0.424</td>
</tr>
<tr>
<td></td>
<td>ZIP Codes</td>
<td>0.362</td>
<td>0.349</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>Census Tracts</td>
<td>0.256</td>
<td><strong>0.295</strong></td>
</tr>
<tr>
<td></td>
<td>ZIP Codes</td>
<td>0.402</td>
<td><strong>0.442</strong></td>
</tr>
<tr>
<td>Boston</td>
<td>Census Tracts</td>
<td>0.028</td>
<td><strong>0.029</strong></td>
</tr>
<tr>
<td></td>
<td>ZIP Codes</td>
<td>0.172</td>
<td><strong>0.185</strong></td>
</tr>
<tr>
<td>Chicago</td>
<td>Census Tracts</td>
<td>0.193</td>
<td><strong>0.262</strong></td>
</tr>
<tr>
<td></td>
<td>ZIP Codes</td>
<td>0.172</td>
<td><strong>0.400</strong></td>
</tr>
<tr>
<td>Houston</td>
<td>Census Tracts</td>
<td>0.255</td>
<td><strong>0.298</strong></td>
</tr>
<tr>
<td></td>
<td>ZIP Codes</td>
<td>0.226</td>
<td><strong>0.327</strong></td>
</tr>
</tbody>
</table>

Table 6: Performance of the cluster-level household income prediction task, in \( R^2 \) scores, when for each test city the model is trained only on the others. We performed also 500 experiments with a null model, reporting the average of the scores for each city.

The choice of the hyperparameter \( q \) is optimized during the cross-validation, between values from 0.1 and 0.5 with step 0.05. In parentheses are shown corresponding percentages of higher-level clusters (census tracts or zonas) for which the prediction is improved respect to the baseline by the regression model. Baseline values are different respect to Table 4 because are evaluated on a different subset of clusters.

5.3. Income Prediction Among Cities

The last experiment is related to the application of a leave-one-out approach for the household income prediction, i.e. using information gained on a training set of multiple cities to estimate the economic variable in a new city that has never been seen by the algorithm. In this section, we apply this method only to the set of US cities, since they share the same aggregation levels. Results are shown for the cluster-level task in Table 6, and for each city is reported the outcome from the model trained on the others. To test the statistical significance of this method, we also report scores when the prediction is performed by randomly assigning cluster-level features among training cities, keeping constant the number of clusters for each city. From the scores discrepancy between our outcomes and the null model we can figure out that our model’s out-of-sample predictive power does not derive from an accidental case.

6. Conclusions

In this work, we investigated the poverty prediction task in the urban environment of two developed countries. We showed that methods used for poverty mapping in resource poor settings can be also applied in this context. Specifically, we have shown that a model pre-trained on the ImageNet dataset can explain, about the target, a significant fraction of the variance with no fine-tuning procedure or proxies. Moreover, we showed that a regression model trained with respect to a given aggregation of the target can infer spatial properties of more granular resolution levels. Finally, we demonstrated the predictive power of these methods if applied to infer the target in new test cities.
Acknowledgments

This project was funded in part by the Data2X Initiative through the Big Data for Gender Challenge Awards. SP acknowledges support from the Lagrange Project and CRT Foundation (http://www.isi.it/en/lagrange-project/project). AP acknowledges support from Intesa Sanpaolo Innovation Center. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References


